UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

COMPUTATIONAL METHODS TO CREATE AND ANALYZE A DIGITAL GENE EXPRESSION ATLAS OF EMBRYO DEVELOPMENT FROM MICROSCOPY IMAGES

TESIS DOCTORAL

Carlos Castro González
Ingeniero de Telecomunicación

2013

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR

DE INGENIEROS DE TELECOMUNICACIÓN

PH.D. THESIS

COMPUTATIONAL METHODS TO CREATE AND ANALYZE A DIGITAL GENE EXPRESSION ATLAS OF EMBRYO DEVELOPMENT FROM MICROSCOPY IMAGES

Author

Carlos Castro González
Telecommunication Engineer

Advisors

Miguel Ángel Luengo Oroz
Telecommunication Engineer, Ph.D.

María Jesús Ledesma Carbayo
Telecommunication Engineer, Ph.D.

March 2013

Biomedical Image Technologies lab
Departamento de Ingeniería Electrónica
Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad Politécnica de Madrid. 2013

Ph.D. Thesis:
Computational Methods To Create And Analyze A Digital Gene Expression Atlas Of Embryo Development From Microscopy Images

Author:
Carlos Castro González

Advisors:
Miguel Ángel Luengo Oroz
María Jesús Ledesma Carbayo

Committee:

D. Andrés Santos Lleó (*presidente*)	*Universidad Politécnica de Madrid, Spain*
D. Arrate Muñoz Barrutia (*secretario*)	*Universidad de Navarra, Spain*
D. Nadine Peyriéras (*vocal*)	*CNRS, Gif-sur-Yvette, France*
D. Thomas Walter (*vocal*)	*Institut Curie, Paris, France*
D. Nicole Gofinkiel (*vocal*)	*CSIC, Madrid, Spain*
D. Emmanuel Faure (*suplente*)	*École Polytechnique, Paris, France*
D. José Luis Rubio Guivernau (*suplente*)	*MedLumics Company, Madrid, Spain*

Edited with LaTeX

Cuando emprendas tu viaje a Ítaca
pide que el camino sea largo,
lleno de aventuras, lleno de experiencias.
No temas a los lestrigones ni a los cíclopes
ni al colérico Poseidón,
seres tales jamás hallarás en tu camino,
si tu pensar es elevado, si selecta
es la emoción que toca tu espíritu y tu cuerpo.
Ni a los lestrigones ni a los cíclopes
ni al salvaje Poseidón encontrarás,
si no los llevas dentro de tu alma,
si no los yergue tu alma ante ti.

Pide que el camino sea largo.
Que muchas sean las mañanas de verano
en que llegues -¡con qué placer y alegría!-
a puertos nunca vistos antes.
Detente en los emporios de Fenicia
y hazte con hermosas mercancías,
nácar y coral, ámbar y ébano
y toda suerte de perfumes sensuales,
cuantos más abundantes perfumes sensuales puedas.
Ve a muchas ciudades egipcias
a aprender, a aprender de sus sabios.

Ten siempre a Ítaca en tu mente.
Llegar allí es tu destino.
Mas no apresures nunca el viaje.
Mejor que dure muchos años
y atracar, viejo ya, en la isla,
enriquecido de cuanto ganaste en el camino
sin aguantar a que Ítaca te enriquezca.

Ítaca te brindó tan hermoso viaje.
Sin ella no habrías emprendido el camino.
Pero no tiene ya nada que darte.

Aunque la halles pobre, Ítaca no te ha engañado.
Así, sabio como te has vuelto, con tanta experiencia,
entenderás ya qué significan las Ítacas.

When you set sail for Ithaca,
wish for the road to be long,
full of adventures, full of knowledge.
The Lestrygonians and the Cyclopes,
an angry Poseidon do not fear.
You will never find such on your path,
if your thoughts remain lofty, and your spirit
and body are touched by a fine emotion.
The Lestrygonians and the Cyclopes,
a savage Poseidon you will not encounter,
if you do not carry them within your spirit,
if your spirit does not place them before you.

Wish for the road to be long.
Many the summer mornings to be when
with what pleasure, what joy
you will enter ports seen for the first time.
Stop at Phoenician markets,
and purchase the fine goods,
nacre and coral, amber and ebony,
and exquisite perfumes of all sorts,
the most delicate fragrances you can find.
To many Egyptian cities you must go,
to learn and learn from the cultivated.

Always keep Ithaca in your mind.
To arrive there is your final destination.
But do not hurry the voyage at all.
It is better for it to last many years,
and when old to rest in the island,
rich with all you have gained on the way,
not expecting Ithaca to offer you wealth.

Ithaca has given you the beautiful journey.
Without her you would not have set out on the road.
Nothing more does she have to give you.

And if you find her poor, Ithaca has not deceived you.
Wise as you have become, with so much experience,
you must already have understood what Ithacas mean.

Ithaca, 1911
Konstantinos P. Kavafis

Acknowledgements

Un viaje finaliza con esta tesis que ha sido mi Ítaca particular durante estos años y a la que no habría llegado sin la la ayuda de muchos compañeros.

Mención especial a mis directores de tesis. A Miguel por iniciarme en el mundo de las células digitales y los sistemas complejos, a María por su vitalidad y optimismo sin fin. Muchas partes de esta tesis (y aquí hablo literalmente) son vuestras. Estos comentarios van también por Andrés que ha seguido este trabajo de cerca desde el principio, me apoyado sin límites en los momentos malos y me ha dado libertad para escoger mis caminos investigadores. Ha sido un placer trabajar con vosotros y espero poder seguir haciéndolo en el futuro.

El viaje ha estado, sin duda, lleno de experiencias y, tal y como pronosticaba Kavafis, me ha llevado a múltiples ciudades extranjeras para aprender de sus sabios. Aquí he de agradecer especialmente la contribución de Paul y Nadine que me acogieron durante mis estancias en París y me regalaron largas y productivas horas de trabajo a su lado, siempre bien recompensadas por un confit de canard o por un sushi en la esquina. Muchas gracias también a Benoît por hacer que rue des Boulets fuera un hogar para mí y por enseñarme la tenacidad necesaria para terminar una tesis. A Thierry, por su hospitalidad sin fin (París no será lo mismo sin las visitas al piso franco de Belleville) y porque es un gran contador de historias, tanto científicas como de las otras. A Louise, porque trabajar a su lado me muestra lo fácil de pasar de lo profesional a lo personal, gracias por tu ayuda y paciencia con mis primerizos pasos en la biología del desarrollo. A Manu, por sus estimulantes charlas sobre *machine learning*. Por último, gracias a Sophie y a Pierre por sus preciosas imágenes y por enseñarme una nueva forma de mirar en ellas. A todos vosotros, y a otros que se me quedarán en el tintero, ¡merci beaucoup!.

Por supuesto, no me puedo olvidar de mis compañeros de barco en el BIT. Tanto los que comenzaron la travesía conmigo (Giorgos, Pedro, Gert,

Juan, Jóse, Ana, Laura, Rosario y Giancarlo), como los que se han ido incorporando después (David, Óscar, Dani, Subhra, Sneha, Suhan), o los que sólo estuvieron un rato (Irene, Ziwei, Evangelina). Gracias por resolverme mis dudas y por el apoyo técnico y el moral. Sin vosotros no hubieran sido igual los viajes a Nueva York, su lista de garitos, las Twin Cities, su "Sevilla", las hamburgesas en medio de Wisconsin, los paseos por Cambridge (Massachusetts), los taxis de noche en Barcelona, las catas de vinos, las pachangas de matados, los proyectos -siempre postpuestos- de triatlones o las múltiples reuniones para festejar tesis, para festejar la navidad o para festejar sin motivo de celebración aparente. Querría mencionar especialmente a Laura con quien he trabajado codo con codo (literalmente) durante toda la tesis, ¡enhorabuena por tu recién adquirida condición de doctora!. También a David, que ha prestado su estrecha colaboración en este trabajo y que prosigue ahora por los apasionantes caminos de la morfogénesis, ¡ánimo con la recta final!.

No quiero dejar en el tintero a mis compañeros de piso, los dos Mikes, que saben bien de las vicisitudes de sacar una tesis adelante y que han sido para mí familia en Madrid. Lo mismo puedo decir de Irene y de Mercedes, de Manu, Rafa y Jara, de Loren y de Santi. Agradecimientos también a Cristina que durante mucho tiempo fue un apoyo y soportó mi tesis. Y a tantos otros que ahora se me olvidan, me falta el tiempo, pero a quienes me encargaré de repetírselo en persona. Gracias por estar ahí y saber que puedo contar siempre con vosotros.

Esta tesis y todo lo que he hecho en la vida se lo debo entero a mi padre y a mi madre. Gracias por escuchar y respaldarme en toda circunstancia. Lo mismo digo de mi hermano Luis y mi hermanita Cristina. Esto va por vosotros.

Carlos Castro González
Madrid, 13 de Marzo de 2013.

Abstract

The creation of atlases, or digital models where information from different subjects can be combined, is a field of increasing interest in biomedical imaging. When a single image does not contain enough information to appropriately describe the organism under study, it is then necessary to acquire images of several individuals, each of them containing complementary data with respect to the rest of the components in the cohort. This approach allows creating digital prototypes, ranging from anatomical atlases of human patients and organs, obtained for instance from Magnetic Resonance Imaging, to gene expression cartographies of embryo development, typically achieved from Light Microscopy.

Within such context, in this PhD Thesis we propose, develop and validate new dedicated image processing methodologies that, based on image registration techniques, bring information from multiple individuals into alignment within a single digital atlas model. We also elaborate a dedicated software visualization platform to explore the resulting wealth of multi-dimensional data and novel analysis algorithms to automatically mine the generated resource in search of biological insights.

In particular, this work focuses on gene expression data from developing zebrafish embryos imaged at the cellular resolution level with Two-Photon Laser Scanning Microscopy. Disposing of quantitative measurements relating multiple gene expressions to cell position and their evolution in time is a fundamental prerequisite to understand embryogenesis multi-scale processes. However, the number of gene expressions that can be simultaneously stained in one acquisition is limited due to optical and labeling constraints. These limitations mo-

tivate the implementation of atlasing strategies that can recreate a virtual gene expression multiplex.

The developed computational tools have been tested in two different scenarios. The first one is the early zebrafish embryogenesis where the resulting atlas constitutes a link between the phenotype and the genotype at the cellular level. The second one is the late zebrafish brain where the resulting atlas allows studies relating gene expression to brain regionalization and neurogenesis. The proposed computational frameworks have been adapted to the requirements of both scenarios, such as the integration of partial views of the embryo into a whole embryo model with cellular resolution or the registration of anatomical traits with deformable transformation models non-dependent on any specific labeling.

The software implementation of the atlas generation tool (Match-IT) and the visualization platform (Atlas-IT) together with the gene expression atlas resources developed in this Thesis are to be made freely available to the scientific community.

Lastly, a novel proof-of-concept experiment integrates for the first time 3D gene expression atlas resources with cell lineages extracted from live embryos, opening up the door to correlate genetic and cellular spatio-temporal dynamics.

Resumen

La creación de atlas, o modelos digitales, donde la información de distintos sujetos puede ser combinada, es un campo de creciente interés en imagen biomédica. Cuando una sola imagen no contiene suficientes datos como para describir apropiadamente el organismo objeto de estudio, se hace necesario adquirir imágenes de varios individuos, cada una de las cuales contiene información complementaria respecto al resto de componentes del grupo. De este modo, es posible crear prototipos digitales, que pueden ir desde atlas anatómicos de órganos y pacientes humanos, adquiridos por ejemplo mediante Resonancia Magnética, hasta cartografías de la expresión genética del desarrollo de embrionario, típicamente adquiridas mediante Microscopía Óptica.

Dentro de este contexto, en esta Tesis Doctoral se introducen, desarrollan y validan nuevos métodos de procesado de imagen que, basándose en técnicas de registro de imagen, son capaces de alinear imágenes y datos provenientes de múltiples individuos en un solo atlas digital. Además, se ha elaborado una plataforma de visualización específicamente diseñada para explorar la gran cantidad de datos, caracterizados por su multi-dimensionalidad, que resulta de estos métodos. Asimismo, se han propuesto novedosos algoritmos de análisis y minería de datos que permiten inspeccionar automáticamente los atlas generados en busca de conclusiones biológicas significativas.

En particular, este trabajo se centra en datos de expresión genética del desarrollo embrionario del pez cebra, adquiridos mediante Microscopía dos fotones con resolución celular. Disponer de medidas cuantitativas que relacionen estas expresiones genéticas con las posiciones celulares y su evolución en el tiempo es un prerrequisito fundamental para comprender los procesos multi-escala característicos de la morfogénesis.

Sin embargo, el número de expresiones genéticos que pueden ser simultáneamente etiquetados en una sola adquisición es reducido debido a limitaciones tanto ópticas como del etiquetado. Estas limitaciones requieren la implementación de estrategias de creación de atlas que puedan recrear un multiplexado virtual de expresiones genéticas.

Las herramientas computacionales desarrolladas han sido validadas en dos escenarios distintos. El primer escenario es el desarrollo embrionario temprano del pez cebra, donde el atlas resultante permite constituir un vínculo, a nivel celular, entre el fenotipo y el genotipo de este organismo modelo. El segundo escenario corresponde a estadíos tardíos del desarrollo del cerebro del pez cebra, donde el atlas resultante permite relacionar expresiones genéticas con la regionalización del cerebro y la formación de neuronas. La plataforma computacional desarrollada ha sido adaptada a los requisitos y retos planteados en ambos escenarios, como la integración, a resolución celular, de vistas parciales dentro de un modelo consistente en un embrión completo, o el alineamiento entre estructuras de referencia anatómica equivalentes, logrado mediante el uso de modelos de transformación deformables que no requieren ningún marcador específico.

Está previsto poner a disposición de la comunidad científica tanto la herramienta de generación de atlas (Match-IT), como su plataforma de visualización (Atlas-IT), así como las bases de datos de expresión genética creadas a partir de estas herramientas.

Por último, dentro de la presente Tesis Doctoral, se ha incluido una prueba conceptual innovadora que permite integrar los mencionados atlas de expresión genética tridimensionales dentro del linaje celular extraído de una adquisición *in vivo* de un embrión. Esta prueba conceptual abre la puerta a la posibilidad de correlar, por primera vez, las dinámicas espacio-temporales de genes y células.

Résumé

La création d'atlas, i.e. de modèles numériques permettant de combiner des informations provenant d'individus différents, est un domaine d'intérêt croissant pour l'imagerie biomédicale. Quand une seule image ne contient pas suffisamment d'information pour décrire complètement l'organisme à étudier, il est nécessaire d'acquérir plusieurs images de plusieurs individus. Chacune de ces images contient alors des données complémentaires par rapport aux autres images. Cette stratégie permet de créer des prototypes numériques comme les atlas anatomiques de patients et organes humains, construits à partir d'images acquises par Résonance Magnétique, ou encore des cartographies d'expressions génétiques, typiquement acquises par Microscopie Optique.

Dans ce cadre, cette thèse propose, développe et valide des nouvelles méthodologies de traitement d'images qui, en se servant des techniques de recalage, alignent des données de plusieurs spécimens sur un seul atlas numérique. Nous avons également développé un logiciel de visualisation spécifiquement dédié à l'exploration de l'énorme quantité de données multidimensionnelles résultant de ces méthodes. De plus, nous proposons de nouveaux algorithmes pour analyser automatiquement les ressources générées en fournissant aux biologistes des interprétations pertinentes.

En particulier, ce travail concentre ses efforts sur des données d'expression génétique acquises l'échèle cellulaire, en microscopie multi-photonique à balayage laser, pendant le développement embryonnaire du poisson zèbre. Disposer des mesures quantitatives qui mettent en relation multiples expressions génétiques avec la position cellulaire et son évolution temporelle est une condition fondamentale pour compren-

dre les processus multi-échelles caractéristiques de la morphogenèse. Toutefois, le nombre des expressions génétiques qui peuvent être acquises au même temps sur la même image est limité par des contraintes expérimentales d'optique et de colorations disponibles. Ces limitations expérimentales motivent l'implémentation de nouvelles stratégies de création d'atlas capables de recréer un multiplex virtuel des expressions génétiques.

Les outils informatiques développés au cours de cette thèse ont été testés sur deux scénarios différents. Le premier concerne l'embryogenèse précoce du poisson zèbre où l'atlas résultant permet d'aborder les liens entre phénotype et génotype au niveau cellulaire. La deuxième concerne le cerveau tardif du poisson zèbre où l'atlas résultant permet alors d'envisager les liens entre expressions génétiques, régionalisation cérébrale, et neurogenèse. Le pipeline déployé a été adapté aux défis particuliers de chacun des scénarios, comme l'inclusion des vues partielles à échelle cellulaire au sein d'un modèle d'embryon entier ou, encore, l'alignement de structures anatomiques de référence en utilisant des modèles de transformations déformables indépendants de une coloration spécifique.

Le software implémenté pour générer les atlas (Match-IT), l'outil de visualisation (Atlas-IT) et les bases de données génétiques générées pendant cette thèse seront disponibles de faon libre pour la communauté scientifique dans le futur immédiat.

Finalement, cette thèse inclut une preuve de concept qui permet d'intégrer des atlas tridimensionnels d'expressions génétiques ainsi que le lignage cellulaire reconstruit à partir de l'imagerie *in vivo* d'un embryon. Cette preuve de concept ouvre la possibilité de corréler pour la première fois les dynamiques spatio-temporelles des cellules avec celles des expressions génétiques.

Contents

To see a world in a grain of sand,
And a heaven in a wild flower,
Hold infinity in the palm of your hand,
And eternity in an hour.

William Blake. *Auguries of Innocence, 1803*

Chapter 1

Motivation and Objectives

1.1 Motivation

Studying how the genetic information relates to the spatiotemporal behavior of cell dynamics, tissue patterning and embryo morphogenesis is one of the key questions for developmental biology in the post-genomic era. Despite having completed the genome sequence of a number of organisms -such as the *C. elegans* worm (Elegans Consortium [1998]), the *Drosophila* fly (Adams et al. [2000]), the mouse (Chinwalla et al. [2002]), the human being (McPherson et al. [2001]), the sea urchin (Sodergren et al. [2006]) or the zebrafish (Ensembl [2007]), we are still far from understanding, modeling and predicting how organisms develop from one single cell into an organized, multicellular individual.

However, comprehensive understanding of biological mechanisms is a fundamental issue for efficient pre-clinical testing of potential new drugs (Goldsmith [2004], Sipes et al. [2011]). Potential applications include treatment of heart diseases (Milan et al. [2003], Barros et al. [2008]), leukemia (North et al. [2007]), bone disorders (Paul et al. [2008]), cancer (Amatruda et al. [2002], Stern and Zon [2003], Lu et al. [2011]), schizophrenia, Parkinson's, Alzheimer's and other dementia (Martone et al. [2008]).

A number of fundamental challenges pave the way towards the long term goal of understanding living systems multiscale dynamics. In order to achieve this goal, it is necessary to quantitatively assess the temporal and spatial gene

expression distribution in multicellular organisms (Lécuyer and Tomancak [2008], Luengo-Oroz [2009]). This is also a requirement for building and modeling gene regulatory networks underlying morphogenesis (Davidson and Erwin [2006], Li and Davidson [2009], Peter et al. [2012]) and to perform stem cell research.

Recent advances in labeling techniques (Vonesch et al. [2006], Choi et al. [2010]) and microscopy imaging (Megason and Fraser [2007]) have steered this field from non-spatial, "omics"-like approaches (Walter et al. [2002]) towards "in-toto" approaches based on images that provide both spatial and temporal quantitative information (Fernandez-Gonzalez et al. [2006], Gorfinkiel et al. [2011]). Hence, the current trend towards automatic, high-content, high-throughput screening brings new bottlenecks in the domain of image analysis (Baker [2010], Truong and Supatto [2011]): The unprecedented rise in complexity and size of data has favored the blossoming of a new discipline, bioimage informatics (Peng [2008], Myers [2012]), or the science of analyzing and organizing massive and heterogeneous biological image data into categorized quantitative information.

In this context, biomedical atlases have emerged as a crucial technology to integrate all this new data deluge into multi-dimensional resources where spatio-temporal changes from the cell to the organism level can be jointly examined Baldock and Burger [2012]. This thesis investigates computational strategies to digitally reconstruct gene expression atlases of vertebrate embryogenesis. The reconstruction of a digital atlas requires a series of image processing steps to map a cohort of individuals onto a common reference space (Fig. 1.1). These operations allow to combine multi-dimensional and multi-subject data and to provide a single representation to visualize, mine, correlate and interpret information at different scales.

The result of this family of methods is the assembly of a digital prototypic model of a "standard" individual which constitutes the essential scaffold where to make the accurate, repeatable, consistent and quantitative measures required for comparative studies (Oates et al. [2009]). Atlases can be compared to geographic information systems (GIS): "spatial databases to which diverse data, primarily but not restricted to imaging data, can be registered and queried" (Martone et al. [2008]). For example, multiple atlas images can be employed to improve segmentation accuracy in medical applications (Artaechevarria et al. [2009]). In

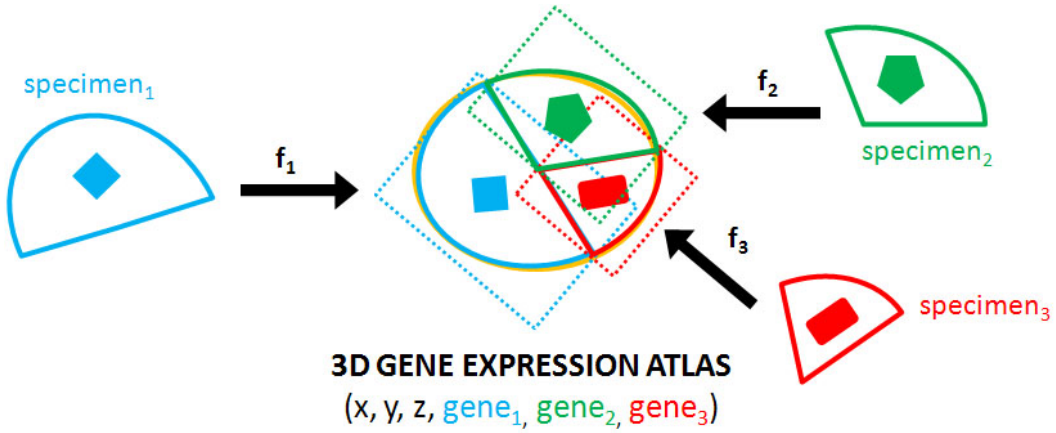

Figure 1.1: This thesis develops computational tools to combine multi-subject, multi-dimensional data into a prototypic, digital model.

addition, atlases are used to identify and categorize anatomic and genetic differences between cohorts of individuals, such as different mutant strains (Warga and Kane [2003]) and constitute an essential tool allowing to relate genotypes and phenotypes (Gorfinkiel et al. [2009]).

In response to the previous motivations, in this work we focus on the digital reconstruction, visualization and analysis of gene expression atlases in zebrafish early and late developmental stages, a problem that has been scarcely treated in the previous literature.

1.2 Objectives

The main objective of this thesis consists in **developing image processing and analysis tools to create an atlas -or digital model- of embryo development that gathers multidimensional gene expression data coming from a cohort of different individuals acquired with state-of-the-art optical microscopy techniques**. These tools will allow to obtain a digital and quantitative description of vertebrate embryogenesis overcoming current limitations in the field (see chapter 2). The main objective is subdivided in three sub-objectives:

1. A framework to reconstruct, visualize and analyze a 3D atlas of gene expression in the early zebrafish embryo

2. A framework to reconstruct a 3D atlas of gene expression in the late zebrafish brain

3. A framework to link gene expression data and cell tracking from early to late zebrafish development

1.2.1 A framework to reconstruct and analyze a 3D atlas of gene expression in the early zebrafish embryo

Current approaches fail to achieve a prototypic model of vertebrates gene expression at the cellular level. This first sub-objective tackles this gap in the state of the art by facing one of the acknowledged challenges in the field (Peng et al. [2011]): the development of registration techniques capable to map high-resolution, but partial images of embryos into one complete template. Achieving this cellular-resolution goal implies developing an image processing chain to detect nuclei centers, segment gene expression patterns and find the corresponding function, $f(e_i, T)$, that registers every partial individual, e_i, into the common template, T. Appropriate visualization tools have to be developed accordingly in order to supervise and validate the whole procedure. The resulting image processing workflow will be applied to early embryos allowing a systematic study of gene co-expression and dynamics. Such workflow has the potential to derive into high content-high throughput image analysis algorithms (Truong and Supatto [2011]).

1.2.2 A framework to reconstruct a 3D atlas of gene expression in the late zebrafish brain

As embryogenesis goes into later stages and cells group into specialized organs, it becomes crucial to develop new mapping software that appropriately includes the apparition of anatomical landmarks into the registration strategy. This sub-objective focus on the zebrafish brain and explores the use of deformable transformation models, such as thin plate splines (Dollár [2006]) and elastic registration schemes (Klein et al. [2009]), to achieve the virtual multiplex of gene expression patterns onto the same target brain.

1.2.3 A framework to link gene expression data and cell tracking from early to late zebrafish development

Although some previous works have tackled the digital reconstruction of cell dynamics (Murray et al. [2008], Keller and Stelzer [2008], Olivier et al. [2010]), the relationship of cell positions and their lineage tree with *in-vivo* gene expressions is still unknown. This integration would require the reconstruction of a spatio-temporal atlas of gene expression which have not been yet achieved because of a number of reasons ranging from labeling to methodological challenges (e.g. availability of transgenic lines and/or temporal synchronization of different specimens' development). This sub-objective makes novel advances in this field by proposing two prospective studies. First, it employs zebrafish transgenic lines labeled to display one gene expression *in vivo* and performs cell tracking to statistically study how gene expression propagates through the lineage tree. Secondly, it proposes a strategy that allows to integrate the 3D gene atlas into a spatio-temporal digital model containing the cell lineage tree. This exploratory research should pave the way to the reconstruction of 4D gene atlas and the understanding of how gene propagation mechanisms are linked to the cell lineage.

1.3 Organization

1.3.1 Operational context

This multidisciplinary PhD thesis has been carried out at Universidad Politécnica de Madrid (Spain) in collaboration with the CNRS and École Polytechnique (France) and involves various different fields ranging from biological procedures and microscopy image acquisition (carried out in collaboration with N&D-CNRS, France), to data processing, quantification and analysis (carried out at BIT-UPM, Spain), or multidimensional data visualization (carried out in collaboration with ISC-École Polytechnique, France). Extensive feedback and interaction, starting with the design of experimental protocols, was carried out between the three partners reflecting on a double iterative workflow, see Fig. 1.2. These collaborations provided a rich training in different transversal subjects through more than

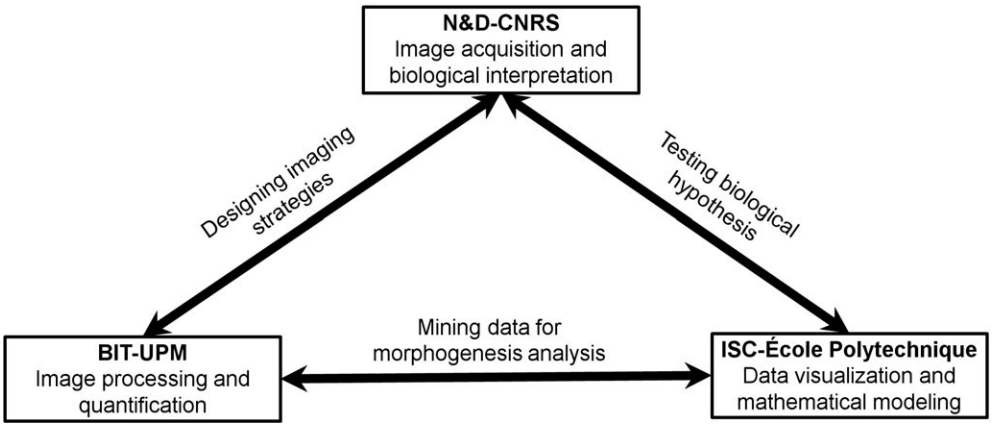

Figure 1.2: Operational context of the present PhD thesis.

8 months of stay in our partner labs, which allows the present thesis to apply for the International PhD mention. The main tasks of each group collaborating during this project are:

- Biomedical Image Technologies (BIT) at Universidad Politécnica de Madrid (UPM) was in charge of image processing and data analysis (Carlos Castro González, Dr. Miguel Ángel Luengo Oroz, David Pastor Escuredo and Evangelina Balanou) under the supervision of Prof. Andrés Santos and Prof. María Jesús Ledesma Carbayo.

- Institut des Systèmes Complexes (ISC) at École Polytechnique, Paris, France was in charge of mathematical modeling with Prof. Paul Bourgine, data visualization with Thierry Savy, human-computer interaction with Dr. Benoit Lombardot and design of the BioEmergences[1] workflow with Dr. Emmanuel Faure and Dr. Camilo Melani.

- Neurobiologie et Developpement (N&D), Institut de Neurobiologie Alfred Fessard, Centre Nationale de la Recherche Scientifique (CNRS), Gif-sur-Ivette, France counted with Dr. Nadine Peyriéras, head of the BioEmergences platform, which was in charge and designing the biological strategy and relevance of the interpretations together with Dr. Louise Duloquin.

[1] http://bioemergences.iscpif.fr

Sophie Desnoulez was in charge of performing fluorescent *in situ* hybridizations of early zebrafish embryos whereas Dr. Pierre Affaticati's technical expertise allowed the acquisition of multiple *in situ* images of the late development in zebrafish brain.

The present PhD thesis was carried out in the Biomedical Image Technologies lab (BIT), DIE, ETSI Telecomunicación at Universidad Politécnica de Madrid (UPM) under the umbrella of the Spanish Ministry of Science and Innovation through the FPU fellowship program, the European projects BIOEMERGENCES (EU-NEST028892) and EMBRYOMICS (EU-NEST012916) and the Spain-France joint scientific action MORPHONET (HF2007-0074).

1.3.2 Document structure

The corpus of this PhD thesis is organized into the following chapters:

1. Motivation and objectives (Chapter 1).
2. State of the art (Chapter 2).
3. A framework to reconstruct, visualize and analyze a 3D atlas of gene expression in the early zebrafish embryo (Chapter 3).
4. A framework to reconstruct a 3D atlas of gene expression in the late zebrafish brain (Chapter 4)
5. A framework to link gene expression data and cell tracking from early to late zebrafish development (Chapter 5)
6. Contributions and future work (Chapter 6).

My mother told me the first things I wrote were continuations of the stories I read because it made me sad when they concluded or because I wanted to change their endings. Writing stories was not easy [...] Fortunately, the masters were there, teachers to learn from and examples to follow.

Mario Vargas Llosa. *Nobel Lecture, 2010*

Chapter 2

State of the art

In this chapter, we critically review the state of the art regarding digital atlases for the anatomy and gene expression of animal organisms. In the following sections, we will discuss in detail the three main fields related to this PhD thesis: Developmental biology, microscopy image acquisition and image processing techniques. In particular, we will deal with the main criteria to take into account when aiming at reconstructing and digitizing embryo development (Table 2.1): Animal models, developmental stages, biological applications, imaging modalities, data types, spatial scales and resolutions, matching procedures, image processing pipelines and visualizations and validation schemes.

The assembly of morphological atlases and the development of appropriate computational strategies have been a major issue in the medical field for years (Park et al. [2003], Aljabar et al. [2009], Fonseca et al. [2011]). For instance, there are several processing and visualization methods devoted to the reconstruction of anatomical human brain atlases (Mazziotta et al. [2001], Toga et al. [2006]). However, the implementation of image processing tools aimed at studying model organisms embryogenesis and gathering an increasing amount of information -including gene expression data- with the sufficiently fine spatial and temporal scales is still an open issue. In this sense, we will describe here the work carried out on a number of model organisms (section 2.1.1), either at the adult stage or throughout embryonic stages (section 2.1.2), imaged with different imaging modalities (section 2.2). We will distinguish atlases focusing either on anatomical structures or gathering genome-wide expression data (section 2.3.1) and limited

either to specific organs (e.g. the brain) or encompassing the whole organism (section 2.3.2). We will also consider different algorithmic strategies that allow to match a population of individuals into one single digital atlas model (section 2.3.3).

2.1 Biological Context

2.1.1 Animal models

Model organisms are chosen for their small size, good properties in terms of phylogenetic position (Fig. 2.2), transparency and/or relevance for studies related to human health.

The nematode *C. elegans*, having the most ancient evolutive emergence among the considered animal models (Fig. 2.2), has a largely invariant cell lineage and stereotyped development which greatly facilitates comparisons between different individuals (Murray et al. [2008], Long et al. [2009], Liu et al. [2009]).

The worm *Platynereis* kept a number of ancestral traits (Tomer et al. [2010]) and proved being insightful for comparative studies.

The fruitfly *Drosophila melanogaster* has been extensively studied in the field of genetics and developmental biology (Fowlkes et al. [2008], Frise et al. [2010], Peng et al. [2011]). 60% of human genetic diseases have their counterpart in the *Drosophila* genome.

The zebrafish (*Danio rerio*) (Fig. 2.3) is more recently a model for developmental biology research thanks to its amenability to genetic investigations and the transparency of its tissues at embryonic stages. In addition, its phylogenetic position closer to human makes it a valuable model for toxicology and pharmacology studies (Hill et al. [2005], Yang et al. [2009]). Producing anatomical and gene expression atlases at different developmental stages is one of the challenges in the field (Castro et al. [2009], Ullmann et al. [2010], Potikanond and Verbeek [2011], Rittscher et al. [2011], Ronneberger et al. [2012]).

Quail (Ruffins et al. [2007]) and chicken (Fisher et al. [2008], Fisher et al. [2011]) are also used as vertebrate models and have interesting features for experimental embryology. The embryo develops outside the egg and is quite well

Reference	Animal model	Imaging modality	Spatial scope and resolution		Developmental period, age* and # of steps			Data type: # of specimens, template composition, # of genes and gene product quantification				Matching type and transformation	
Murray et al. 2008	C. elegans	*In vivo* microscopy	Whole organism	Cellular	Early	4-350 cells	5h	20	Individual	4	YES	Object-based	Non-rigid
Long et al. 2009	C. elegans	Microscopy	Whole organism	Cellular	Early	L1	1	15	Individual	0	NO	Object-based	Affine
Liu et al. 2009	C. elegans	Microscopy	Whole organism	Cellular	Early	L1	1	324	Synthetic	93	YES	Object-based	Non-rigid
Tomer et al. 2010	Platynereis	Microscopy	Brain	Cellular	Late	48h	1	171	Average	8	NO	Intensity-based	Affine + non-rigid
Fowlkes et al. 2008	Drosophila	Microscopy	Whole organism	Cellular	Early	0-100%	6	1822	Average	95	YES	Object-based	Non-rigid
Frise et al 2010⁺	Drosophila	Microscopy	Whole organism	Multicellular	Early	5%	1	2693	Synthetic	1881	YES	Object-based	Non-rigid
Peng et al. 2011	Drosophila	Microscopy	Brain	Tissular	Adulthood	Adult	1	2945	Average	470	NO	Object-based	Affine + non-rigid
Potikanond et al. 2011	Zebrafish	Microscopy	Whole organism	Tissular	Late	24-72h	4	≈75	Individual	≈20	NO	Semantic	-
Rittscher et al. 2011	Zebrafish	Microscopy	Whole organism	Organular	Late	120h	1	11	Average	0	NO	Object-based	Rigid + non-rigid
Ullman et al. 2010	Zebrafish	MRI	Brain	Tissular	Late	4 months	1	1*	Individual	0	NO	None	None
Ronneberger et al. 2012	Zebrafish	Microscopy	Brain	Tissular	Late	48,72,96h	3	85	Average	17	NO	Object-based	Non-rigid
Castro et al. 2009	Zebrafish	Microscopy	Whole organism	Cellular	Early	6h	1	6	Individual	5	YES	Intensity-based	Rigid
Ruffins et al. 2007	Quail	MRI	Whole organism	Organular	Late	5-10d	6	6*	Individual	0	NO	None	None
Fisher et al. 2011	Chicken	OPT	Wing bud	Tissular	Late	3-5d	3	45	Individual	5	NO	Object-based	Non-rigid
Baldock et al. 2003⁺	Mouse	OPT	Whole organism	Organular	Late	2-17d	22	23484	Individual	14927	NO	Object-based	Non-rigid
Carson et al. 2005	Mouse	Microscopy	Brain	Cellular	Late	7d	1	200	Synthetic	200	YES	Object-based	Affine + non-rigid
Kovacevic et al. 2005	Mouse	MRI	Brain	Tissular	Late	56d	1	9	Average	0	NO	Intensity-based	Affine + non-rigid
Ma et al. 2005	Mouse	MRI	Brain	Tissular	Late	56d	1	10	Probabilistic	0	NO	Intensity-based	Affine + non-rigid
Johnson et al. 2010	Mouse	MRI	Brain	Tissular	Late	56d	1	14	Probabilistic	0	NO	Intensity-based	Affine
Lein et al. 2007⁺	Mouse	Microscopy + MRI	Brain	Cellular	Late	56d	1	≈20000	Average	≈20000	YES	Intensity-based	Rigid + non-rigid
Woods et al. 1999	Human	MRI	Brain	Organular	Adulthood	Adult	1	22	Average	0	NO	Intensity-based	Non-rigid
Rex et al. 2003	Human	MRI	Brain	Tissular	Adulthood	Young adult	1	452	Average	0	NO	Intensity-based	Affine + non-rigid
Smith et al. 2004	Human	MRI	Brain	Tissular	Adulthood	Adult	1	58	Average	0	NO	Intensity-based	Affine
Kerwin et al. 2010	Human	OPT	Brain	Tissular	Late	47-55d	2	5	Individual	3	NO	Object-based	Non-rigid

* h=hours, d=days, % of membrane invagination, ⁺ 2D gene expressions, * No matching procedure: Each specimen was directly employed as the template of its corresponding developmental stage.

Table 2.1: Overview of recent strategies for reconstructing digital anatomy and gene expression atlases for animal organisms. Adapted from Castro-González et al. [2012].

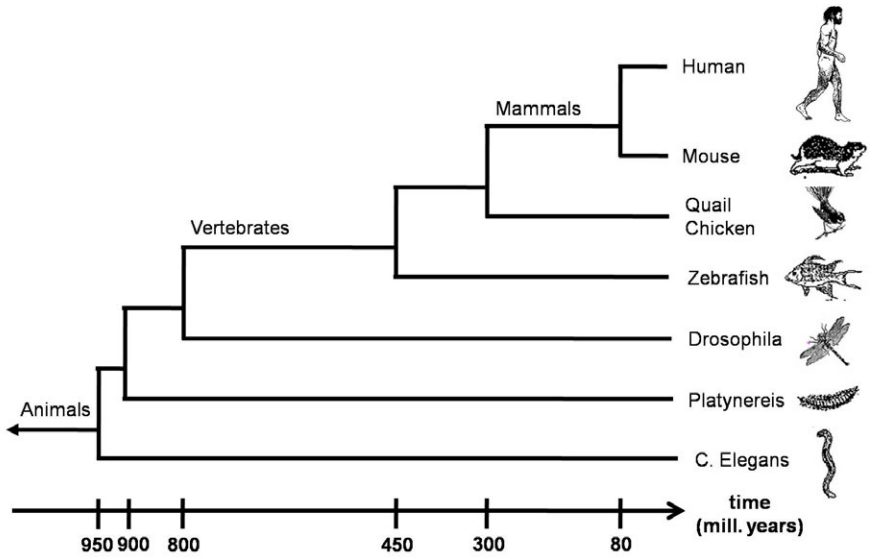

Figure 2.2: Timing of evolutionary emergence and phylogenetic relationships of different model organisms. Time estimations were extracted thanks to the tool developed by Hedges et al. [2006].

amenable to *in vivo* imaging (section 2.2) and the construction of atlases at the cellular level (see section 2.3.2).

Mouse is the major mammalian model organism for biomedical investigations and much effort has been devoted to the reconstruction of their development (MacKenzie-Graham et al. [2004], Ma et al. [2005], Kovačević et al. [2005], Carson et al. [2005], Lein et al. [2007], Johnson et al. [2010], Richardson et al. [2010], Hawrylycz et al. [2011]). The relatively large size of the mouse embryo makes difficult to capture the whole specimen in a single-shot, *in-toto* imaging strategy (section 2.2) with sufficient spatial resolution (section 2.3.2).

The same difficulty applies to fixed human embryos (Woods et al. [1999], Rex et al. [2003], Smith et al. [2004], Kerwin et al. [2010]) where the creation of a standard cartography of human brains (Hawrylycz et al. [2012]) is of fundamental importance in medical studies.

2.1.2 Developmental Stages

The construction of anatomical and gene expression atlases has attracted scientific attention during the whole embryogenesis process, and the proposed methodolo-

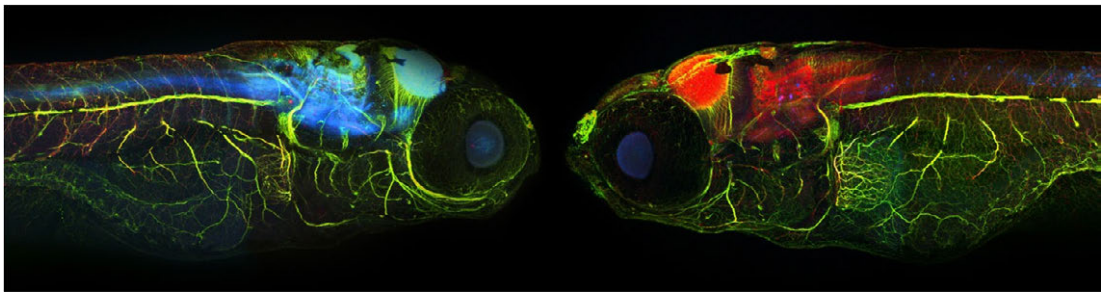

Figure 2.3: Confocal microscopy acquisitions of zebrafish, combined with fluorescent stains, can be employed to study genetic Alzheimer's *in vivo*. In this image, the nerve cells are stained in green while the Alzheimer's genes are colored in blue and red (Paquet et al. [2009]).

gies depend on the developmental stage analyzed which may vary from early stages to adulthood (Table 2.1). At early developmental stages, the whole organism is more easily amenable to *in toto* imaging with resolution at the single cell level, Fig. 2.4D. This is the case for animal models such as the *C. elegans*, the *Drosophila* or the zebrafish (Fowlkes et al. [2008], Castro et al. [2009], Long et al. [2009]). At later developmental stages or in the adult, it can be more relevant to focus on specific organs (Fig. 2.4.A-C) such as the brain in humans, mice, *Drosophila* or zebrafish (Woods et al. [1999], Lein et al. [2007], Peng et al. [2011], Ronneberger et al. [2012]) or the appendages in the chicken (Fisher et al. [2011]).

Most studies targeted a single developmental stage (Fig. 2.4.G). However, the temporal scale is essential for the understanding of biological mechanisms and gathering atlases with the relevant kinetics is a major trend in the field (Fig. 2.4.E). Fisher et al. [2011] reconstructed atlases combining fate mapping data and gene expression patterns for 3 consecutive developmental stages of the chick wing bud. Murray et al. [2008] took advantage of *C. elegans* largely invariant lineage to build the first 3D+time atlas of transgenic reporters expression patterns in *C. elegans* from the 4-cell stage to the 350-cell stage.

2.1.3 Biological Applications

The digital reconstruction of anatomical and gene expression atlases paves the way towards new biological insights (Luengo-Oroz et al. [2011]). Below, we enumerate some of the most prominent results and applications derived from these

atlases.

Kovačević et al. [2005] used an atlas model to perform genetic and anatomic phenotyping, achieving the automated detection of mutant strains. Atlases also had major implications in evolutionary studies and Tomer et al. [2010] identified related parts of the brain in phylogenetically distant animals. Chiang et al. [2011] created a comprehensive brain wiring map of the adult *Drosophila* brain which provides a crucial tool to analyze information processing within and between neurons.

Using all the genetic information gathered in their model, Frise et al. [2010] clustered genes co-expression domains to elucidate previously unknown genetic functions and molecular and genetic interactions. Lein et al. [2007] detected highly specific cellular markers and deciphered cellular heterogeneity previously unidentified in the adult mouse brain with a gene expression atlas of more than $20,000$ genes. Carson et al. [2005] discovered gene expression patterns in the mouse which are possibly related to Parkinson disease.

Liu et al. [2009] showed in *C. elegans* that different gene regulatory pathways can lead to identical cell fates. Cell fate modules with specific molecular signatures repeatedly occurred along the cell lineage revealing bifurcations towards cell differentiation. By computationally analyzing gene expressions in chick embryos, Fisher et al. [2011] identified the cell fate leading to digits formation. The association of gene expression atlases with cell tracking techniques and fate maps generation (McMahon et al. [2008], Pastor et al. [2009], Supatto et al. [2009]) are promising tools for stem cell studies and regenerative medicine.

2.2 Image Acquisition Background

Three main imaging modalities have been employed in literature for the assembly of digital atlases: Fluorescence microscopy, magnetic resonance imaging and optical projection tomography.

Each of these modalities have different resolutions and lead to different types of atlases (Table 2.1). The choice depends on the specimen thickness and its optical transparency. For each animal model (section 2.1.1) these properties vary with the age of the specimen (section 2.1.2).

Recent advances in photonic microscopy imaging (Fig. 2.4.A) include multi-harmonic (Evanko [2010]) and fluorescence imaging by confocal, multi-photon laser scanning microscopy (Abbott [2009], Pardo-Martin et al. [2010], Mahou et al. [2012]) or light-sheet fluorescence microscopy (LSFM) (Huisken and Stainier [2009], Keller et al. [2010], Tomer et al. [2012]). When combined with newly developed fluorescent proteins, biological sensors (Chudakov et al. [2005], Giepmans et al. [2006]) and in-situ hybridization (ISH) techniques (Welten et al. [2006], Brend and Holley [2009]) microscopy techniques open new perspectives for the construction of high resolution anatomical and gene expression atlases (Conrad et al. [2011]). Spatial resolution of hundreds of nanometers and temporal resolution of minutes have been achieved for the observation of entire organisms at different levels of organization. However, photonic microscopy imaging is still limited to small model organisms with good optical properties.

Optical projection tomography (OPT) (Sharpe et al. [2002]) was introduced as an alternative optical method to fluorescence microscopy and overcomes the limitation of the specimen thickness. OPT generates data by acquiring many views of the same specimen at different rotation angles then assembled to create a 3D volume (Fig. 2.4.C). OPT resolution in the range of millimeters does not however allow working at the cellular level.

Alternatively, magnetic resonance imaging (MRI) (Jacobs et al. [2003]) does not use fluorescent staining and has thus a broad range of applications (Fig. 2.4.B). Indeed, MRI contrast does not depend on the penetration of photons but on the voxel-to-voxel variations in water content leading to diverging spins when being excited by magnetic fields. MRI achieves a spatial resolution of about tens of microns. Although more and more intense magnetic fields are used, single cell resolution is still far from the state of the art.

2.3 Image Processing For Gene Atlas Reconstruction

2.3.1 Data Types

A number of different specimens is assembled in the constructions of atlases models that can just carry anatomical information or multilevel, genome-wide data.

Anatomical atlases providing a scaffold with the morphological and histological landmarks characteristic of a cohort (Rex et al. [2003], Ruffins et al. [2007], Ullmann et al. [2010]) constitute a reference shape or template to integrate further information coming from other individuals and reflect the intrinsic variability of morphogenesis processes.

Genome-wide atlases (Fisher et al. [2008], Richardson et al. [2010]) integrate gene expression patterns and multilevel information from various sources into anatomical atlases (Fig. 2.4.D). This approach emulates a virtual multiplexing and overcomes the restrictions in the number of gene products and/or functional patterns that can be simultaneously assessed. As a consequence, they are becoming a major tool for making spatiotemporal correlations between the different levels of biological organization, comparing individuals, building prototypic models and deciphering the relationship between genotypes and phenotypes.

Building an anatomical atlas requires defining a common scaffold, frequently called template, where to gather all the information collected from different specimens. There are diverging criteria in the literature about how an atlas template should be built. Several studies (Ruffins et al. [2007], Castro et al. [2009], Ullmann et al. [2010]) employ one single individual to match all the rest of the population (Fig. 2.4.D). This individual is chosen for its "standard" appearance and the corresponding data should be of the highest quality. Alternatively, an iterative method has been used to identify the median individual within a population and select it as the template (Long et al. [2009]). Using an individual as template has the advantage of allowing to merge the atlas information with the cell lineage tree (see chapter 5). Other works (Frise et al. [2010]) used a synthetic template to map all the data from a cohort of specimens. This template consists in an engineered "virtual specimen" which retains the essential features of a species. The use of an

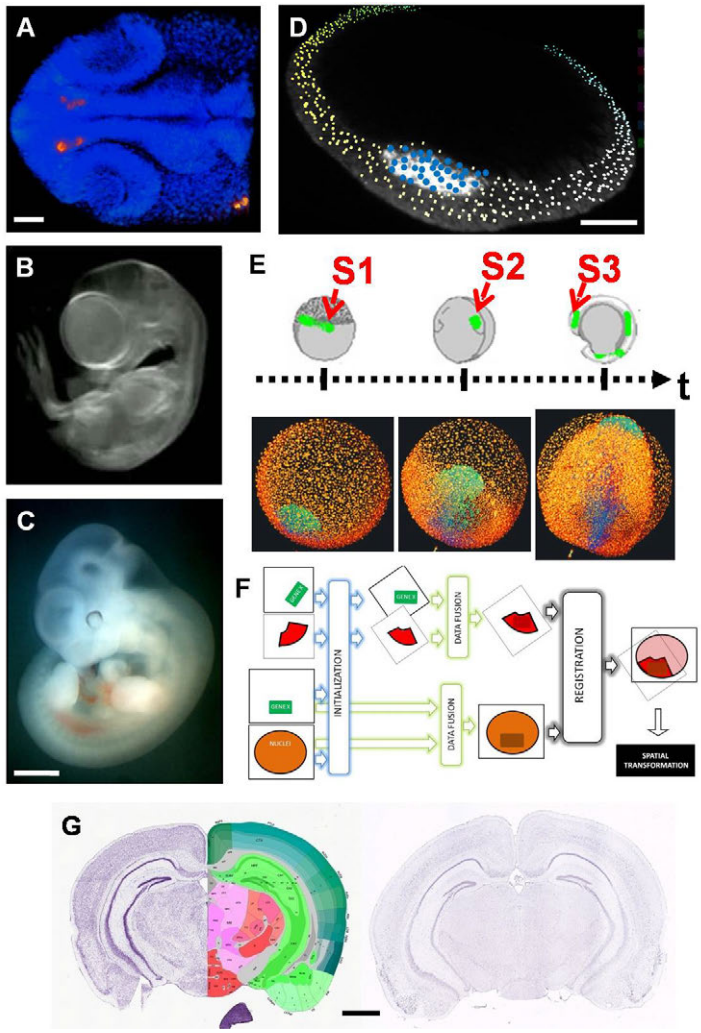

Figure 2.4: Examples of components involved in an atlas model construction. **A:** Confocal microscopy acquisition of a 24h zebrafish brain labeled by fluorescent ISH (Thyrosine hydroxylase RNA probe) and DAPI staining of cell nuclei. Scale bar 100 microns. **B:** MRI of a quail extracted from Caltech's "Quail Developmental Atlas". Available from: http://131.215.15.121/. **C:** OPT of a late mouse embryo extracted from the "EMAP eMouse Atlas Project", http://www.emouseatlas.org. Scale bar 1000 microns. **D:** Orthoslice showing the nuclei of an early zebrafish model where the raw gene expression from another specimen has been integrated. Cells positive for the expression of the gene are highlighted in blue. Scale bar 100 microns. **E:** Early zebrafish templates for three different developmental stages where individuals can be mapped according to a reference gene pattern. **F:** Reconstruction of a mosaic-like atlas: Guided by a reference pattern, partial views of different individuals are mapped into a complete template. **G:** *Left panel*, Coronal section of an averaged 3D template showing organ-level anatomical annotations of an adult mouse brain at a given developmental stage. *Right panel*, an ISH slice warped into the atlas template through deformable transformation models. Extracted from the "Allen Mouse Brain Atlas [Internet]. Seattle (WA): Allen Institute for Brain Science. 2009". Available from: http://mouse.brain-map.org. Scale bar 1300 microns.

average template (Fig. 2.4.G) is widely spread (Rex et al. [2003], Fowlkes et al. [2008], Peng et al. [2011]). Ma et al. [2005] constructed a "minimal deformation average template" as an idealized specimen minimizing the deformation required to fit any specimen of the cohort. Although average templates usually exhibit a better definition and a higher signal-to-noise ratio in regions very similar among specimens, they fail to faithfully model fine features and regions with a high variability, lowering their definition (Kovačević et al. [2005], Dorr et al. [2008]). Finally, some approaches used a probabilistic template (Johnson et al. [2010]) where specimens variability is represented by statistical confidence limits.

The construction of prototypical genome-wide atlases implies imaging gene expression patterns in 3D with resolution at the cellular level (Hendriks et al. [2006]). Image processing methods are required to achieve the automated segmentation of gene expression domains and the quantification of gene products to allow for example the description of expression domain borders. A simple quantification strategy is based on the assumption of a linear relationship between fluorescence intensity and gene expression level (Wu and Pollard [2005], Frise et al. [2010]). The obtained measurements are often normalized with respect to the nuclei channel fluorescence (Liu et al. [2009]), which is considered to be constant, to compensate for thickness-dependent signal detection. Normalization with respect to the background intensity (Murray et al. [2008]) is also a common strategy. Another possibility consists in clustering a population of cells into discrete levels (e.g. strong, moderate, weak and none) depending on the gene expression signal intensity (Carson et al. [2005]). Although the three different methods yielded correlated measurements across different individuals, the relevance of the obtained quantitative measurements to compare different specimens is questionable and this issue remains a challenge. Current efforts to achieve the quantitative comparison of gene expression levels in different individuals include the minimization of variability within a population (Fowlkes et al. [2008]) and the conversion of fluorescence signal into fluorescent protein numbers in transgenic specimens (Damle et al. [2006]).

2.3.2 Spatial scales and resolution

Constructing a prototypic model for an organism can achieve different scopes, from particular organs to the whole organism, which can be resolved at either the organular, tissular, multicellular and eventually cellular resolutions (Table 2.1).

The reconstruction of atlases with resolution at the cellular level (Fig. 2.4.D) generally focused on the species more phylogenetically distant from human (section 2.1.1) and generally addressed early developmental stages (section 2.1.2). Identifying every single cell position in the whole imaged specimen (Long et al. [2009]) requires advanced image processing methods (section 2.3.4). Algorithmic strategies for the approximate detection of the cell nuclei center in 3D volumes encompassing several thousands of cells have been described (Frolkovic et al. [2007], Drblikova et al. [2007], Krivá et al. [2010]). In addition, the identification of cell contours helps assessing RNAs or protein expression to the cell. Voronoi geometries have been proposed as a simple approach to determine cellular boundaries (Luengo-Oroz et al. [2008]). The real cell shape can be obtained by the algorithmic segmentation of cell membranes in 3D when the latter is available (Malpica et al. [1997], Zanella et al. [2010], Mikula et al. [2011], Mosaliganti et al. [2012]).

Working at the mesoscopic scale -e.g. with multicellular neighborhoods- is less demanding and already provides useful information (Fisher et al. [2008], Frise et al. [2010]).

Annotating and segmenting the different anatomical structures of interest at the tissular level is required to reconstruct prototypic models of organs. Examples of such methods can be found in Ma et al. [2005], Kovačević et al. [2005], Dorr et al. [2008], Johnson et al. [2010] and Ullmann et al. [2010].

Finally, large organisms with a huge number of cells and high complexity in terms of organization led to organ-based atlases that do not resolve the single cell level (Fig. 2.4.G). This strategy has been used for vertebrates at late developmental stages (section 2.1.2) when the specimens size and lack of optical transparency does not allow imaging with resolution at the single cell level (Baldock et al. [2003], Ruffins et al. [2007], Rittscher et al. [2011]).

2.3.3 Matching Procedures

To create an atlas, a matching procedure is required to import each specimen (the source) into the template according to the maximization of a likelihood criteria. Repeating this operation is the core of digital atlases construction. For the same purpose, medical imaging makes extensive use of registration techniques (Maintz and Viergever [1998], Zitova and Flusser [2003]). Three main registration techniques to build digital atlases can be distinguished according to the information used to assemble the data and the minimization criteria chosen accordingly: Intensity-based, object-based and semantic-based registrations. We can also distinguish three different transformation types between the source and the template space: Rigid, affine or non-rigid.

Prior to registration, an initialization scheme is generally applied in order to get a rough alignment between source and template. The initialization scheme helps the registration to reach an accurate solution. Two common initialization techniques consist on coarsely aligning anatomical landmarks (Lein et al. [2007]) or the major orientation axis of an organism such as the anterior-posterior or dorsal-ventral axis (Blanchoud et al. [2010], Balanou [2010]). Qu and Peng [2010] developed an original skeleton standardization technique to rule out part of the geometrical variability between *Drosophila* embryos. In the same line, Peng et al. [2008] designed a method to straighten *C. elegans* worms into the same canonical space.

The populations of individuals to be registered are normally composed of complete specimens imaged similarly. Accurately matching cohorts of partial specimens (Fig. 2.4.F) is one of the current challenges in the field (Peng et al. [2011]) and very few strategies have addressed this case (Castro et al. [2009]).

2.3.3.1 Transformation categories

Rigid transformations are applied when the mapping between the source and template spaces consists on spatial translations and rotations (Castro et al. [2009]). Rigid registration has the advantage of keeping the original raw data unaltered allowing faithful measurements and validation of the true volumes in the final atlas representation. Affine transformations (Rex et al. [2003], Smith et al. [2004])

also include scaling factors apart from translations and rotations. Both rigid and affine transformations are linear and globally applied to all voxels.

On the contrary, non-rigid transformations are non-linear and locally warp the source image to fit into the template (Woods et al. [1998], Ng et al. [2007], Ng et al. [2009], Rittscher et al. [2010]). This typically results in an alteration of the original raw data. For instance, non-rigid warping has been widely used to reconstruct 3D volumes from serially sectioned 2D images in electron microscopy (Saalfeld et al. [2012]) and in series of histological sections, which are often heavily distorted due to folding of the tissues (Arganda-Carreras et al. [2010]).

2.3.3.2 Intensity-based registration

Intensity-based registration procedures align the source and template by maximizing a similarity metric (typically mutual information or cross-correlation) between the gray level values in the voxels of both images.

The most common approaches (Lein et al. [2007], Tomer et al. [2010], Asadulina et al. [2012]) include an initialization performed by a global, intensity-based affine or rigid registration, followed by local deformable warps (Fig. 2.4.G). Multi-resolution approaches are also employed to optimize the mapping procedure in a coarse-to-fine strategy (Smith et al. [2004], Kovačević et al. [2005], Tomer et al. [2010]). Finally, multi-modal approaches combine information coming from different imaging modalities, merging, for instance, histology and MRI (MacKenzie-Graham et al. [2004], Johnson et al. [2010]). Such approaches provide multiple entry points to match different individuals and heterogeneous populations into the same coordinate system.

2.3.3.3 Object-based registration

Object-based transformations attempt to bring into alignment equivalent sets of characteristic points or landmarks present in both the source and template images (Liu et al. [2009]). These transformations are local and non-linear and typically produce an alteration of the data shapes and volumes.

Peng et al. [2011] and Ronneberger et al. [2012] separately developed automatic pattern recognition systems to identify visual anatomic references with

certain geometric properties such as high local curvature. These landmarks are typically employed to guide Thin Plate Splines registration methods (Mace et al. [2006]). Another approach, successfully applied to *C. elegans* embryos (Fowlkes et al. [2008]), consisted in creating methods that unequivocally identify cell to cell correspondences between the template and the individual embryos.

2.3.3.4 Semantic-based registration

Unlike intensity-based and object-based registrations, semantic-based registration does not operate on the geometrical space but on the use on semantical information coming from standard ontotologies (Ashburner et al. [2000]) and web queries (Zaslavsky et al. [2004], Potikanond and Verbeek [2011]).

After following an annotation procedure for anatomy and gene expression data with a controlled, standard vocabulary, the mapping procedure is reduced to just linking names to positions or domains (Baldock et al. [2003], Boline et al. [2008]). Given the difficulty of geometric registration across greatly variable resources, this strategy is useful to guarantee inter-operability and can bring together data coming from different laboratories, resources, developmental stages or even different species.

2.3.4 Image Processing Pipeline

The construction of atlases, or digital representations of anatomic and genetic features from an increasing amount of more and more complex data, requires sophisticated image analysis algorithms (Khairy and Keller [2011]) replacing non-efficient and time consuming processing performed manually or through generic imaging software.

All the image processing steps required to gather quantitative, genome-wide data from a cohort of individuals in a prototype with resolution at the cellular level can be conceptualized in a generalized processing pipeline (Fig. 2.5). This pipeline can achieve a complete (x, y, z, G) model including quantitative data for gene or protein expression level (G) in each cell position (x, y, z) in a developing model organism (Castro et al. [2011]).

This process can involve preprocessing steps such as image enhancement

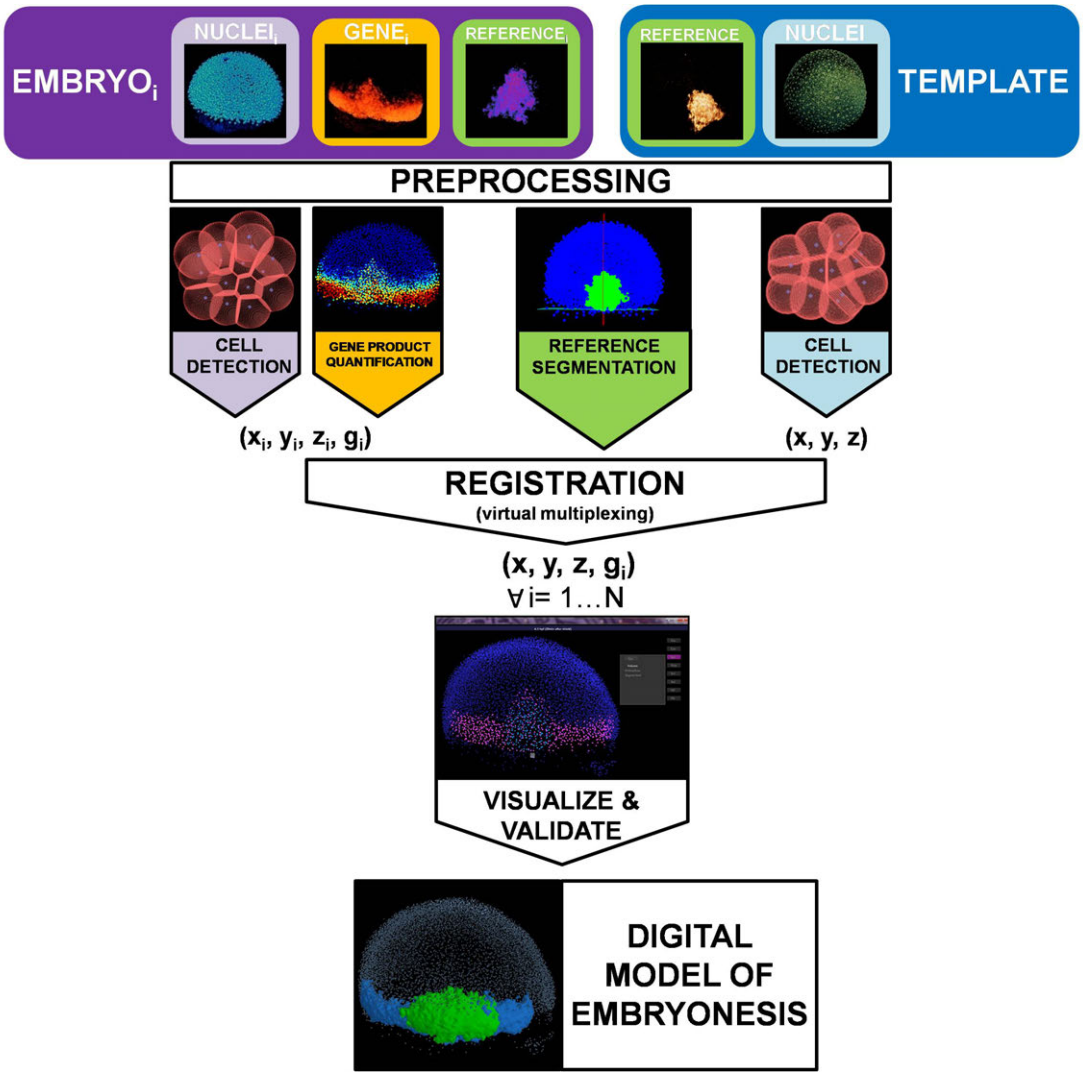

Figure 2.5: Image processing pipeline to build a digital atlas model: After preprocessing, nuclei detection and cell segmentation algorithms are applied to extract cell position and volume. This information is combined with quantification schemes. These operations, iterated throughout a cohort of individuals, yield cellular-level, quantitative measurements of a number of genetic and/or functional patterns. A common reference and/or landmarks highlighted in all individuals are automatically segmented and identified to steer a registration procedure. The latter multiplex all measured patterns into a single, digital template. The resulting atlas can be validated and mined through dedicated, interactive visualization tools.

or multi-view fusion algorithms (Sorzano et al. [2009], Rubio-Guivernau et al. [2012]). Then, cell nuclei detection and cell segmentation techniques (section

2.3.2) are applied to extract cells position, (x_i, y_i, z_i), and volume. Next, signal quantification procedures (section 2.3.1) are applied on gene expression or functional patterns to identify positive cells, (x_i, y_i, z_i, g_i). Repeating this procedure for all the individuals of a cohort yields measurements for relevant patterns, $(x_1, y_1, z_1, g_1)...(x_N, y_N, z_N, g_N)$, which are finally combined through a registration procedure (section 2.3.3). The anatomical information extracted during the cell detection step and automatically segmented or identified landmarks guide the registration process. The final result is a single, quantitative model of the specimen development, $(x, y, z, g_1...g_N)$. Validation of the model and further analysis should use a dedicated, custom-made, interactive visualization interface (section 2.3.5).

2.3.5 Visualization and Validation

The reconstruction of digital atlases relies on automatic algorithms that can handle the enormous amount of large 3D images providing multilevel data for cohorts of individuals at different developmental stages. The lack of gold standards in the field requires the manual curation and correction of the results (Long et al. [2009]).

Several indirect validation techniques have been exploited: Fowlkes et al. [2008] and Peng et al. [2011] showed that the gene expression variability in their atlas model was comparable to that shown by individuals, implying that the experimental errors introduced in the model could be considered negligible. Fisher et al. [2011] applied hierarchical clustering (Pearson correlation) to several gene expression copies coming from different specimens and demonstrated that they grouped together as expected.

In addition to these indirect validation measures, visual assessment is the common validation standard for virtually all the previously described strategies (Table 2.1). Consequently, a number of sophisticated visualization platforms has been developed to display the multi-dimensional input data and output results and to interactively run the previously described methods on request while providing the necessary tools to correct, annotate, quantify and mine their outcomes (Long et al. [2012]). These platforms represent the necessary trade-off between

the automated, high-throughput, fast computer algorithms and the manual, low-throughput but accurate human interactions.

A comprehensive review of such visualization tools can be found in Walter et al. [2010] or Eliceiri et al. [2012]. Some relevant instances include: FlyEx (Pisarev et al. [2009]), GoFigure (Gouaillard et al. [2007]), PointCloudExplore (Weber et al. [2009], Rübel et al. [2010]), Mov-IT (Olivier et al. [2010]), BrainExplorer (Lau et al. [2008]), BrainGazer (Bruckner et al. [2009]), CellProfiler (Jones et al. [2008]) and Vaa3D (Peng et al. [2010]).

> Every move the puzzler makes, the puzzlemaker has made before; every piece the puzzler picks up, and picks up again, and studies and strokes, every combination he tries, and tries a second time, every blunder and every insight, each hope and each discouragement have all been designed, calculated, and decided by the other.
>
> GEORGE PEREC. *Life A User's Manual, 1978*

Chapter 3

A framework to reconstruct, visualize and analyze a 3D atlas of gene expression in the early zebrafish embryo

3.1 Introduction

Deciphering and integrating genetic and cellular dynamics underlying living systems morphogenesis and homeostasis is a major challenge of the post-genomic era. Although full genome sequence is available for a number of animal model organisms (Schier and Talbot [2005]), quantitative data for the spatial and temporal expression of genes is lacking (Oates et al. [2009]).

Progress in photonic microscopy imaging (Megason and Fraser [2007], Abbott [2009], Supatto et al. [2011]) combined with remarkable advances in labeling techniques (Chudakov et al. [2005], Vonesch et al. [2006]) allowed gathering data at all levels of living systems organization with adequate spatial and temporal resolutions. Fluorescent *in situ* hybridization techniques (Brend and Holley [2009], Choi et al. [2010]) immunocytochemistry and transgenesis, combined with 3D optical sectioning, allow assessing gene expression dynamics throughout animal development with precision at the cellular level. However, moving forward from

2D databases of gene expression -such as those available in most model organisms- to multidimensional atlases is a major paradigm shift still requiring the development of appropriate methods and tools.

In this context, developing automated strategies for building gene expression atlases with resolution at the cellular level is a major bottleneck towards biological insights, requiring new methodological and engineering solutions (Luengo-Oroz et al. [2011], Khairy and Keller [2011]). Constructing prototypes for cohorts of individuals from imaging data is a necessary step of atlases construction. This can be solved by finding a spatial correspondence between individuals based on registration methods, a technique used in medical imaging (Maintz and Viergever [1998], Zitova and Flusser [2003]). However, gathering in a prototype multimodal and multiscale features from different specimens with phenotypic variability (Hendriks et al. [2006]) remains a challenge.

Recent studies on different model organisms explored computational strategies for building atlases either assessing cell position to build prototypic specimens (Long et al. [2009], Olivier et al. [2010]) or gathering gene expression patterns observed in cohorts of specimens (Lein et al. [2007], Fowlkes et al. [2008], Peng et al. [2011]) and very few frameworks combined both features. Long et al. [2009] gathered data from 15 *C. elegans* specimens at the earliest larval stage (L1 with 357 cells) to build a statistical 3D atlas of nuclei positions. *C. elegans* presents a number of advantages facilitating the reconstruction process. The entire organism can be imaged with resolution at the single cell level and its cell lineage tree is stereotyped enough to allow the spatiotemporal matching of different individuals at the single cell level. The same feature allowed the construction of a prototypic lineage for a cohort of 6 specimens of *Danio rerio* (zebrafish) embryos throughout their first 10 cell divisions (Olivier et al. [2010]). Peng et al. [2011] achieved the spatial matching of 2,945 adult *Drosophila* brains to gather the expression patterns of 470 different genes. Similarly, Lein et al. [2007] constructed a comprehensive atlas of the adult mouse brain containing about 20,000 gene patterns. The first gene expression atlas with resolution at the cellular level was constructed by Fowlkes et al. [2008] integrating 95 gene expression patterns observed at 6 different developmental stages in a total of 1,822 different *Drosophila* embryos within a common 3D stencil. Applying this approach to ver-

tebrate model organisms faces other difficulties from higher cell lineage variability to gene expression patterns heterogeneity and highly dynamic features at the spatial and temporal level. In addition, the construction of gene expression pattern atlases for vertebrates implied overcoming one of the acknowledged methodological challenges in the field (Peng et al. [2011]): the acquisition of partial volumes taken at high resolution from single specimens and their precise mapping into complete reference specimens. Building 3D atlases of gene expression for the zebrafish blastula and gastrula stages brings specific difficulties due to the lack of morphological landmarks required for achieving patterns registration (Fowlkes et al. [2011], Ronneberger et al. [2012]). The zebrafish, a vertebrate model organism increasingly used for its relevance for biomedical applications (Stern and Zon [2003]), gathers good properties for investigating the reconstruction of its early embryogenesis multiscale dynamics.

The Gene Regulatory Network (GRN) architecture (Fig. 3.1) for the zebrafish early embryonic development is currently under construction (Chan et al. [2009], Longabaugh et al. [2009]) and the data from a 3D atlas of gene expression with resolution at the cellular level is expected to provide the necessary measurements for further modeling the GRN dynamics (Jaeger et al. [2004], Crombach et al. [2012b], Peter et al. [2012]) and possibly integrating the genetic and cellular level of organization (Delile [2013]). Such data would make the zebrafish the first vertebrate model amenable to such a systemic study.

We provide here a methodology to construct, visualize and analyze gene expression atlases for vertebrate early developmental stages. We designed, implemented and deliver here Match-IT and Atlas IT, computational frameworks to automatically map, visualize and analyze 3D gene expression patterns from different individuals (the *analyzed embryos*), onto a common reference specimen (the *template*), with resolution at the cellular scale. The resulting virtual "multiplexing procedure" overcomes the limited number of gene products that can be jointly stained and analyzed in a single specimen.

Match-IT (section3.3) was used to produce the prototypic cartography of 9 gene expression patterns imaged in 3D from double fluorescent in situ hybridization at 6 developmental stages (see section3.5). Atlas-IT (section3.4) proved being suitable for the interactive visualization of gene co-expression patterns and

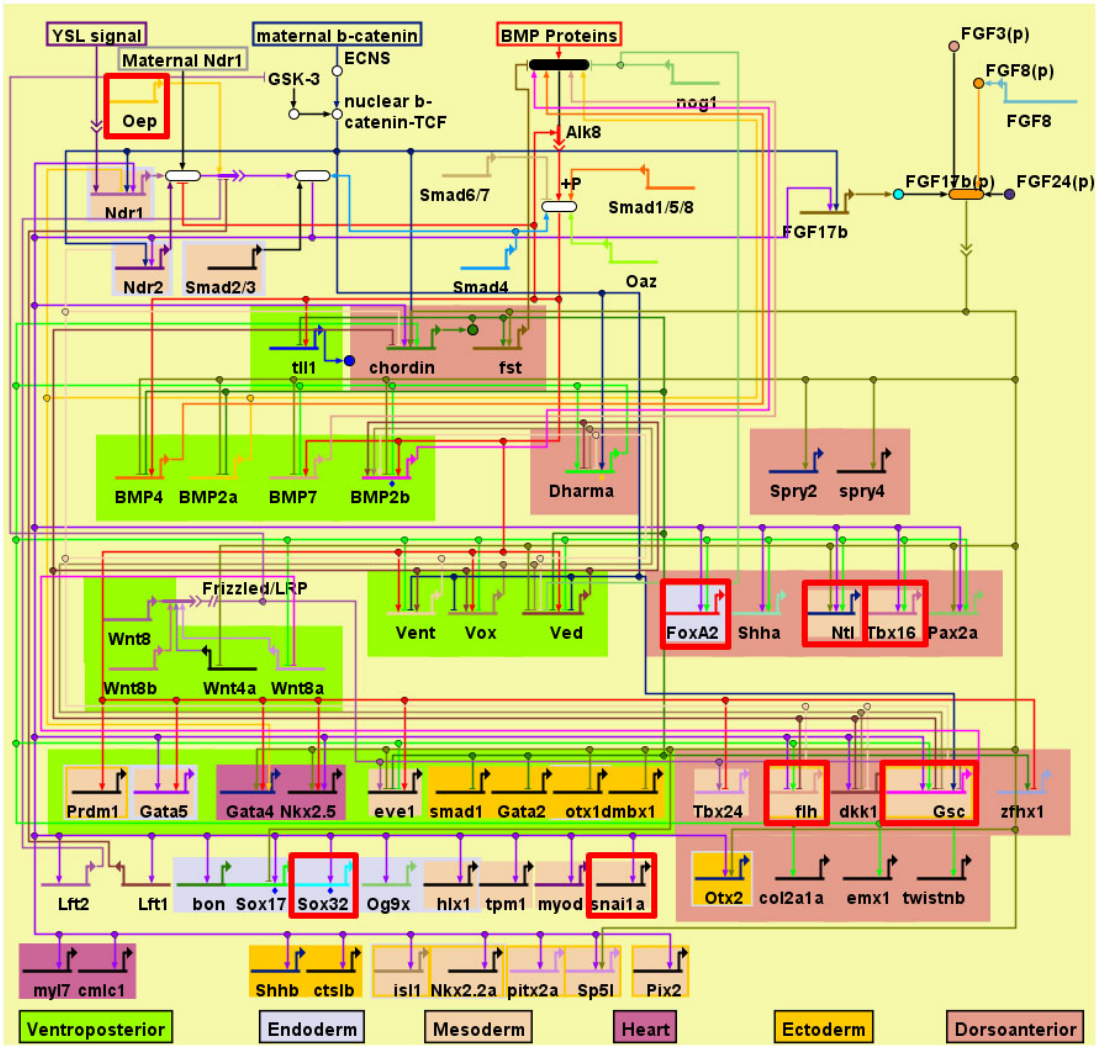

Figure 3.1: The gene expression atlas introduced in this chapter (see section 3.5) includes several gene expression products (in red squares) that play an important role in the zebrafish gene regulatory network (GRN) designed by Yuh's lab in biotapestry (Chan et al. [2009]).

their dynamics. We validated (section3.6) our 3D atlas construction methodology with the automated and quantitative assessment of gene patterns similarity and overlap through time. Analytical tools (section 3.7), such as clustering, were designed to identify morphogenetic domains and genes synexpression groups. The zebrafish blastula and early gastrula atlases with resolution at the cellular scale provide means for the integration of genetic and cellular data unavailable so far.

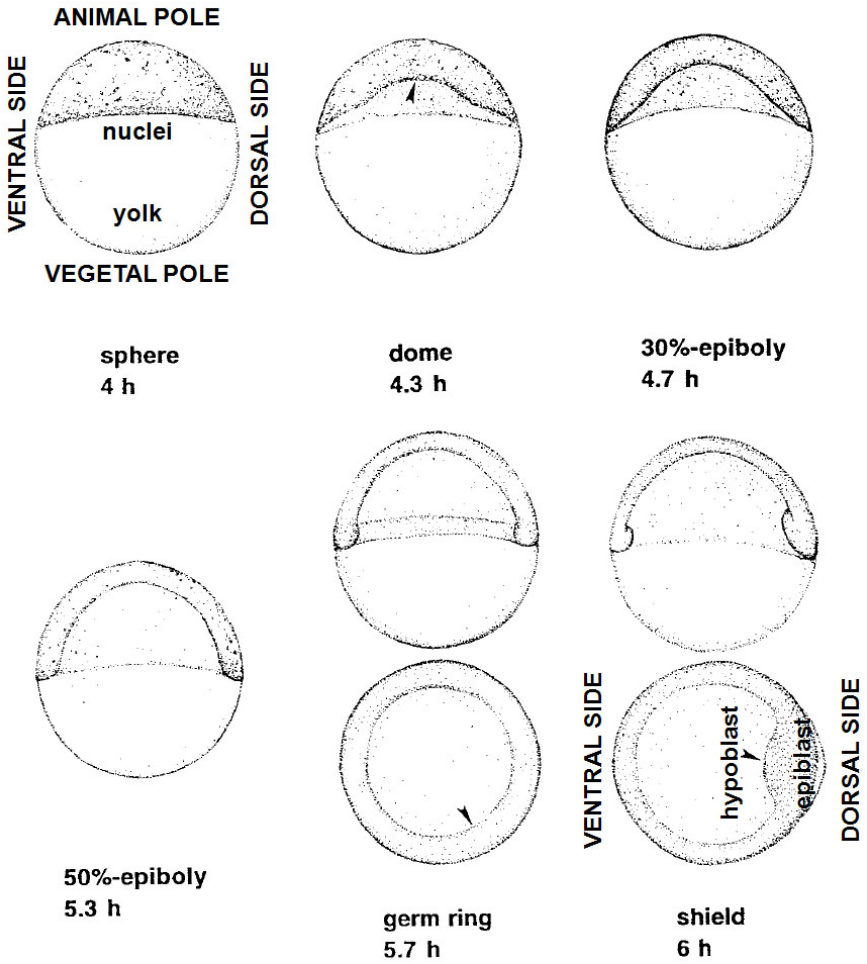

Figure 3.2: Zebrafish embryo development between 4 and 6 hpf. The main anatomical references during these developmental stages are the animal and vegetal poles and the dorsal and ventral sides. Adapted from Kimmel et al. [1995].

3.2 Data acquisition

Biological data consisted in 3D images acquired by confocal laser scanning microscopy from zebrafish embryos fixed at early development, particularly, at 6 different time points: *sphere* (4 hpf), *dome* (4.3 hpf), *30% epiboly* (4.7 hpf), *50% epiboly* (5.3 hpf), *shield* (6 hpf) and *late shield* (6.3 hpf) according to the staging defined at 28.5° C (Kimmel et al. [1995]), see Fig. 3.2. This developmental period is of special interest as precursor of gastrulation (at 6 hpf) when cells begin to migrate along the dorsal midline leading to the formation of two differ-

entiated cell layers: hypoblast (inner) and epiblast (outer), see Fig. 3.2. These images were acquired by confocal laser scanning microscopy from fixed zebrafish embryos, with fluorescent staining of gene expression patterns (Brend and Holley [2009]) and DAPI counterstained to highlight the cell nuclei which were further processed for nuclear center detection (see section 3.3.2). The overall framework aimed to provide the best compromise for embryos staining, data acquisition, nuclei detection and patterns segmentation to achieve the mapping procedure and the selection of gene expression positive cells in the atlas template (Fig. 3.3).

The atlas *template* was based on the 3D volume of an entire specimen with (1) highlighted nuclei and (2) stained *gsc* expression domain, imaged at each of the 6 chosen developmental stages. Nuclei and *gsc* expression domains were also systematically revealed in all the *analyzed embryos*, serving as references to later compute the mappings from the *analyzed embryos* to the *template*. In addition to *gsc* staining, each *analyzed embryo* was processed for double fluorescent *in situ* hybridization staining the pattern of another gene of interest among the 8 additional chosen expressions: *sox32*, *tbx16*, *oep*, *snai1a*, *foxa2*, *ntla*, *flh* and *egfp*. This second spatial expression pattern is later to be mapped on the *template* once the mapping has been computed using the nuclei and *gsc* data. The *template* data was obtained by confocal laser scanning microscopy imaging through the whole embryo with a 10x objective while the *analyzed embryos* were imaged with a high numerical aperture 20x objective providing a 3D view limited to the dorsal side of the embryo to achieve a better resolution (Fig. 3.5a).

3.2.1 *In situ* hybridization and confocal imaging

In vitro fertilization was used to synchronize the spawn. Embryos, staged according to Kimmel et al. [1995], were fixed 24h at 4°C in PFA 4% then rinsed 3 times in PBS 0.1% Tween and stored at −20°C in ethanol. Double fluorescent in situ hybridization (FISH) was carried out as described in Brend and Holley [2009] using antisens RNA probes coupled to fluoresceine or digoxigenine. Probes were detected with an anti-digoxigenin-POD Fab fragment and a anti-fluoresceine-POD Fab fragment (Roche) used at 1:250 dilution in the blocking reagent. Revelation of probes were done with tyramides, as substracts of POD, which were coupled

to fluorescent proteins (Cy3 or Cy5 mono NHS ester, from Amersham ; NHS Fluoresceine, from Pierce) according to the protocol of P. Vize (Zhou and Vize [2004]). Embryo nuclei staining were performed by bathing embryos 1h at room temperature in DAPI (In vitro gen D3571) diluted in PBS 1/1000. Image acquisition was performed with a Leica SP2 bi-photon and confocal laser scanning microscope equipped with a Leica objective HCX APO 20X/0,5W U-V-I or HCX APO 10X/0,3. Embryos were mounted in teflon mold, oriented in a dorsal view, and maintained thanks to 1% agarose perfusion.

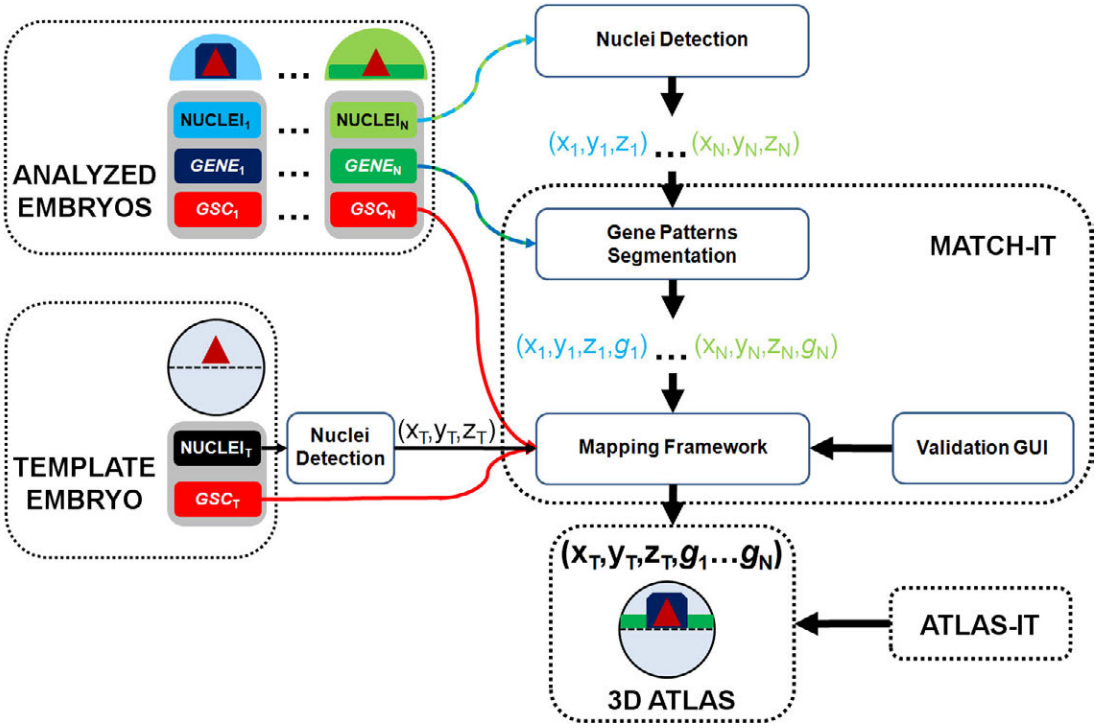

Figure 3.3: Schematic illustration of the atlas construction workflow. For each developmental stage, the *analyzed embryos*, which are partial 3D volumes of an entire specimen, undergo nuclei detection and gene patterns segmentation before being mapped into one common *template*. The mapping procedure is guided by the specimen's shape given by nuclei staining and a common reference gene expression pattern encoding the *goosecoid* (*gsc*) transcription factor. Each step in the workflow can be visualized and inspected for validation with both Match-IT and Atlas-IT. The final model, where all the gene patterns coming from different individuals can be jointly compared, constitutes a 3D atlas.

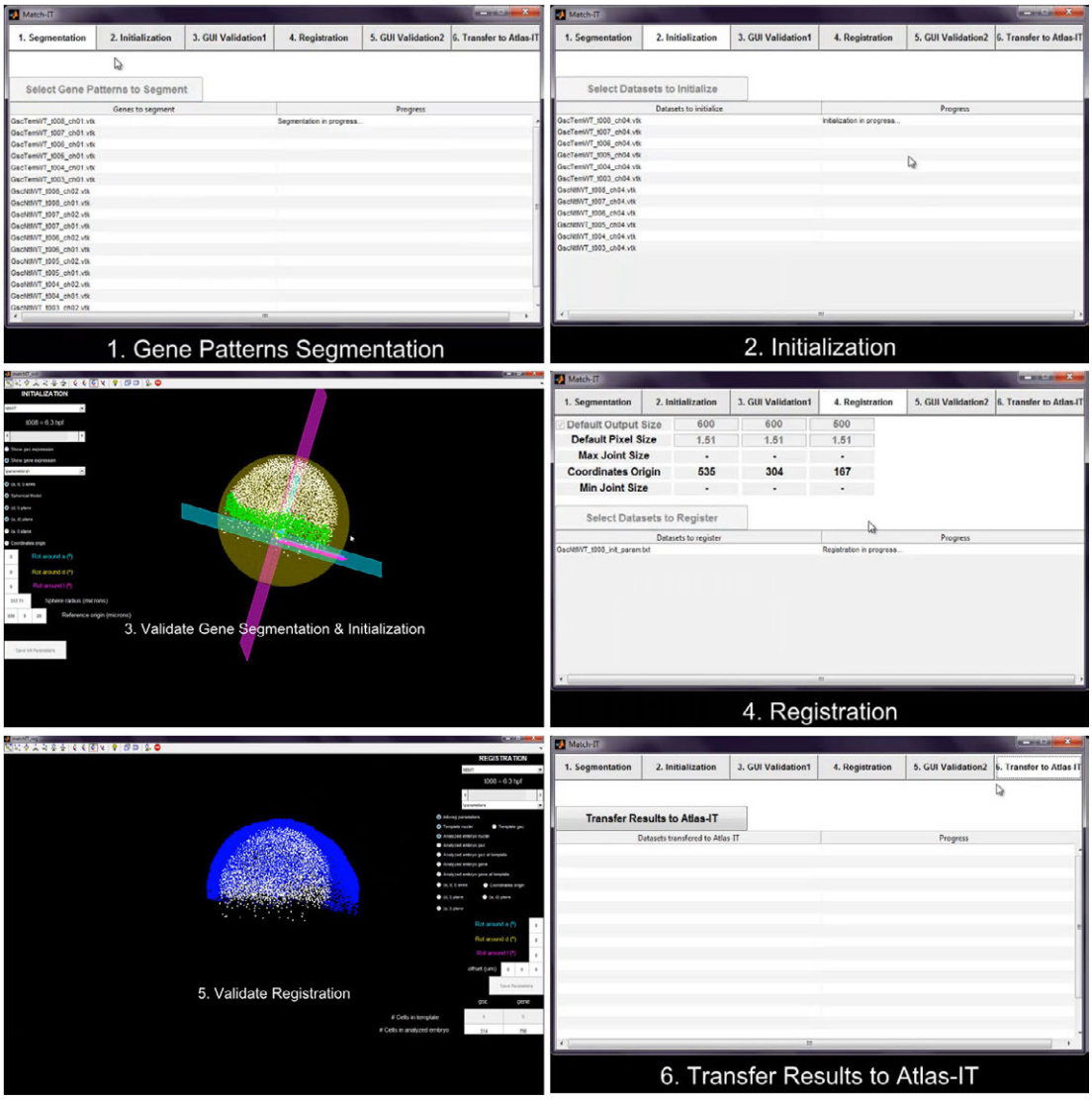

Figure 3.4: Step-by-step procedure to map an *analyzed embryo* into the atlas *template* using Match-IT.

3.3 Match-IT: A workflow to build a 3D atlas of gene expression

3.3.1 Problem overview

We designed and deliver a generic computational framework, going from image acquisition to image data analysis to perform the mapping of different stained

gene expression patterns onto a common prototypic model to create a 3D atlas of gene expression with resolution at the cellular scale. We designed the methodology to be generic for animal model organisms at early developmental stages. The atlas was planned to be precise, easy and efficient to query and analyze, and readily available for the integration of more 3D data.

The atlas construction process consisted in finding the appropriate transformation to match the partial 3D volumes of the *analyzed embryos* within the corresponding *template* (Fig. 3.5a-b). To achieve this goal, the original software Match-IT performed the algorithmic segmentation of gene expression domains, the mapping of *analyzed individuals* onto a common reference specimen and identified positive cells to finally deliver a 3D atlas summarizing single cells genetic profile (Fig. 3.3, Fig. 3.4).

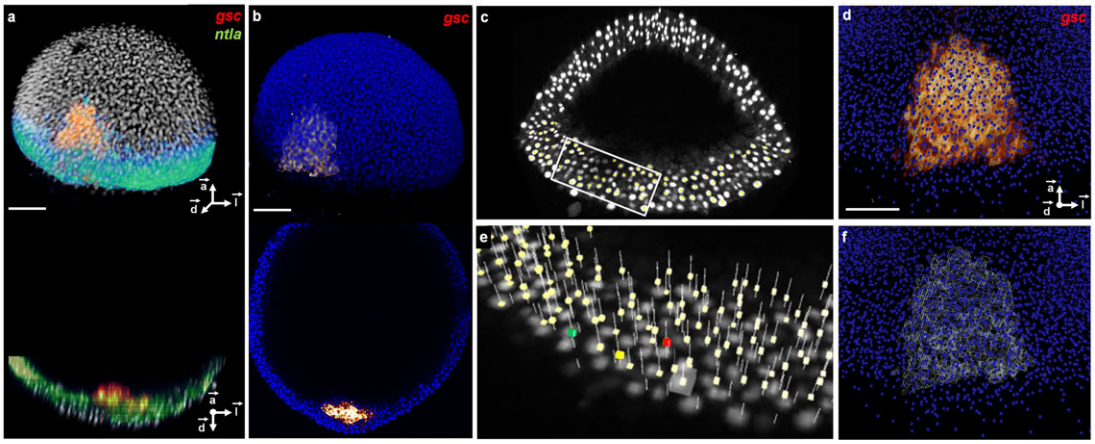

Figure 3.5: 3D raw data, nuclei detection, gene patterns segmentation and their validation at 6.3 hpf. (**a**) *Upper panel:* Volume rendering and *lower panel:* axial orthoslice of an *analyzed embryo* nuclei (in white), the reference gene pattern *gsc* (in red) and *ntla* pattern (in green). (**b**) Same as (a) for the *template* with nuclei in blue and *gsc* pattern in red. (**c**) Nuclei positions in yellow superimposed to raw data in white displayed with three orthoslices (augmented phenomenology). (**e**) 3X zoom on the boxed region in (c) with an example of validated nucleus in green, false positive in red, false negative in yellow, the white cube indicates a selected position to be evaluated. (**d**) Zoom on the *template gsc* raw expression (in red) superimposed to the *template* nuclei positions (in blue). (**f**) Same as (d) with the segmented *gsc* domain in white. Scale bars, 100 μm. Axes point to the animal pole ($\vec{a}$), dorsal side ($\vec{d}$) and lateral side ($\vec{l}$) of the embryo respectively.

3.3.2 Nuclei detection

Precise nuclei center detection served to assess the embryonic axes and define a referential common to all the observed specimens, to provide the delineation of the embryo contours taken as morphological reference, and to assign positive gene expression at the single cell level.

Nuclear detection was achieved by an automated algorithmic strategy with supervision of the chosen parameter set by the expert and validation or, when necessary, correction by visual inspection (Fig. 3.5c,e). For zebrafish data, less than 3% detection errors were estimated from manual validation. This error rate was considered acceptable to further proceed to gene expression domains segmentation and their mapping onto the *template*.

In particular, nuclei centers were detected as local maxima of a smoothed, simplified version of the original image. The preprocessing was performed first by convolving the image with two Gaussians with two different standard deviations in the range of 1.5-2.5 μm and 12-16 μm respectively. Then calculating their difference and preserving only gray values greater than a given threshold that could vary between the 1 and 10%. This allowed to simultaneously smooth the image and preserve only significant objects. Multiple simulations were run combining different possible values for the Gaussians, standard deviations and threshold. Optimal values were visually chosen by interactively checking the detection results with the custom designed visualization software, Mov-IT (Olivier et al. [2010]).

Based on this supervised detection of cell nuclei, we performed a statistical estimate of the distance between neighboring nuclei in each embryo. This distance was taken as the typical cell diameter. This parameter decreased as a function of time as expected through cell divisions (see Fig. 3.6). We used this information to build a cell mask around each nucleus and capture the overall specimen's shape. The former was used to assign gene expression to a single cell and the latter guided the registration procedure to match the *analyzed embryos* into the *template*.

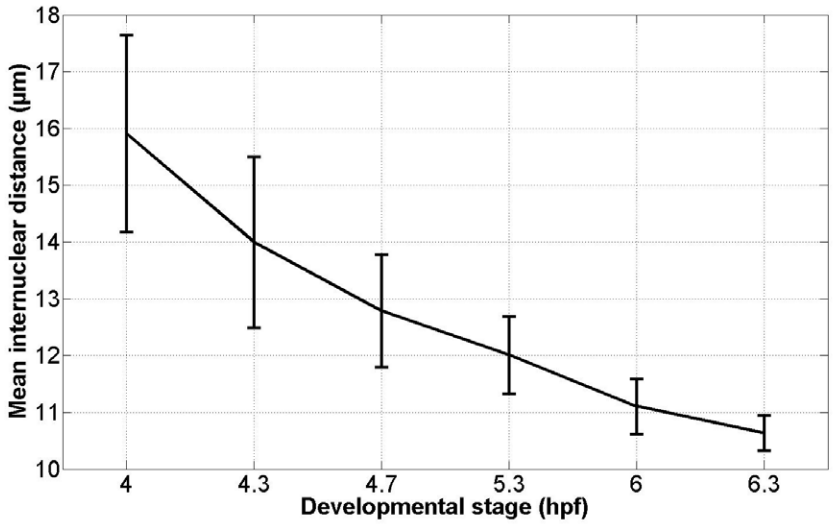

Figure 3.6: Evolution of the internuclear distance through time. The mean internuclear distance (in μm) is calculated for 9 different specimens at each stage. The observed decrease fits with an average of approximately 1.75 divisions per cell between 4 and 6.3 *hpf* and an exponential decrease in the cell cycle length. This is in agreement with previous observations (Kimmel et al. [1995], Keller and Stelzer [2008]) and validates the accuracy of the nuclei detection procedure. Standard deviation is interpreted to reflect individual variation.

3.3.3 Supervised Segmentation of the Gene Expression Patterns

Segmenting the gene expression domains first required the supervised choice of the lower image intensity values that best defined the domain features. Match-IT used these parameters to perform a thresholding operation followed by a morphological closing and area opening (Serra [1982,1988]), leading to the segmentation of a single connected domain with no holes. The result was validated by inspection with Atlas-IT (Fig. 3.5d,f) and allowed selecting positive cells either in the *analyzed embryo* or in the *template* after the mapping operation.

3.3.4 Positive cells selection

Cells in the *analyzed embryos* were selected as positive for the expression of a given gene if their nuclei approximate centers were less than half the average internu-

clear distance from the domain's border. Identifying in the analyzed specimens cells positive for genes expression allowed us to compare domains co-stained in the same *analyzed embryo* to domains gathered in the atlas *template* but stained in different *analyzed embryos*, as shown in Fig. 3.9c-d.

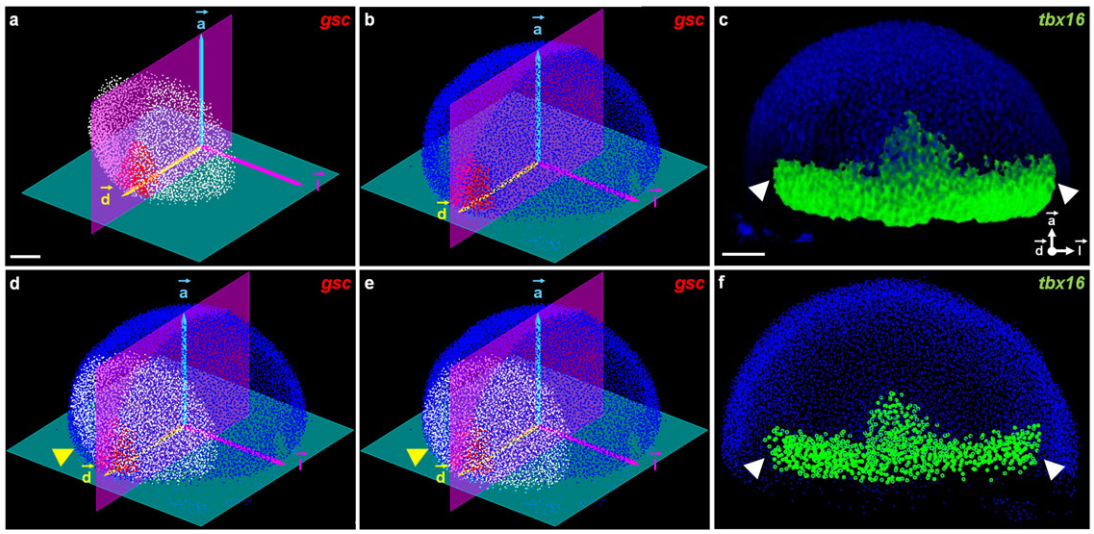

Figure 3.7: The mapping procedure for the 6.3 hpf atlas. (**a**) *Analyzed embryos*, detected nuclei in white, *gsc* positive cells in red, plane passing through the blastoderm margin (in green), bilateral symmetry plane (in purple) and the referential ($\vec{a}$, $\vec{d}$, $\vec{l}$) automatically extracted by the initialization scheme. (**b**) Same as (a) for the *template*, detected nuclei in blue. (**d**) Initialization step aligning the ($\vec{a}$, $\vec{d}$, $\vec{l}$) basis for the *analyzed embryo* and the *template*; yellow arrowhead points to an initialization mismatch refined in (**e**) through the intensity-based registration procedure. (**c**) *tbx16* raw gene expression pattern (in green) from the *analyzed embryo* mapped into the *template* raw nuclei (in blue). (**f**) Detected *template* nuclei (in blue) falling into the analyzed gene expression domain are considered positive (in green). White arrowheads indicate the limits of the imaged *analyzed embryo*. Scale bars 100 μm.

3.3.5 Mapping framework

Mapping the partial 3D volumes of the *analyzed embryos* on the corresponding *template* consisted in finding the appropriate transformation matching the embryos' common references, their outer cell layer contours and their *gsc* positive domains, Fig. 3.7a,e. The mapping procedure was a two-steps process. First, we performed a coarse initialization based on the extraction of a common referential

in the different specimens, Fig. 3.7a,b. Then, a pixel-based registration was used to refine the alignment, Fig. 3.7d,e.

3.3.5.1 Reference detection

Because the zebrafish early embryos largely lacked distinctive anatomical landmarks, the mapping initialization process was based on the automated identification of a common referential, Fig. 3.7a,b. The latter was defined by (1) the plane P_1 separating the cells from the yolk at the embryo margin and (2) the bilateral symmetry plane P_2, computed as the plane orthogonal to P_1 and containing both the center of the embryo spherical approximation and the center of mass of the *gsc* positive nuclei population. These planes unequivocally defined a three-vector basis comprising the animal-vegetal axis ($\vec{a}$), the dorso-ventral axis ($\vec{d}$) and the perpendicular vector given by the right-handed trihedron ($\vec{l}$). The origin of this triplet and the triplet itself were used to compute a translation and a rotation respectively to find equivalent points in different embryos. The result of this matching initialization was supervised and, if necessary, manually corrected with the Match-IT graphical user interface, developed to minimize the effort of manual supervision, Fig. 3.7a,b,d,e.

In particular, this initial reference extraction starts by performing a spherical fit to the detected cell nuclei in all *analyzed embryos* and *templates*. The embryo margin results from a plane fit to the 5 % southernmost nuclei. The bilateral symmetry plane results from connecting the spherical model center and the center of mass of the *gsc* segmentation perpendicularly to the blastoderm margin. The origin of the triplet is placed at the latitude of the blastoderm margin and the longitude defined by the center of mass of the *gsc* segmentation.

3.3.5.2 Grayscale registration

This coarse initialization step was refined by a pixel-based registration procedure. Because the zebrafish early embryos largely lack distinctive morphological references to apply landmark-based registration methods (Fowlkes et al. [2008], Tomer et al. [2010], Peng et al. [2011]) and given the partial nature of the volumes to be aligned, we opted for a rigid, pixel-based transformation scheme that

searched the optimal translation and rotation to match the specimens' volumes (Castro et al. [2009]), see Fig. 3.7d,e. The rigid transformation preserved original gene patterns and the possibility to go back to the raw data for visualization/validation/correction with the Match-IT software at every step of the processing pipeline including the final registration tuning. At this final step, the mean manual corrections to refine the mapping of *analyzed embryos* onto the *template* were in the range of 13 μm of offset, i.e. approximately one cell row, and 3 degrees of rotation per embryo. A Graphical User Interface (GUI) was developed to minimize the effort required for manual validation and correction (Fig. 3.7e).

3.3.6 Cell-to-cell assignation

Finally, the same rule described for the selection of positive cells in the *analyzed embryos* was used to determine positive *template* cells (Fig. 3.7c,f). The number N_i of cells positive for the expression of the gene i in the *analyzed embryo* differed from the number M_i of cells identified as positive in the *template* after the mapping procedure. The difference was mainly due to individual variation in terms of internuclear distance and consistent with an uncertainty of one cell row in the mapping procedure (see Fig. 3.8).

3.4 Atlas-IT: A dedicated tool for the visualization of 3D atlases of gene expression

The analysis of the 3D atlas required dedicated visualization tools to test hypotheses and derive biological insights. Available software did not fulfill our requirements. Either too specific for other model organisms -e.g. PointCloudXplore for *Drosophila* (Rübel et al. [2010])- or limited to other applications -e.g. AceTree for cell lineage (Boyle et al. [2006])- or designed more as general visualization/processing tools -Vaa3D (Peng et al. [2010]), GoFigure (Gouaillard et al. [2007]), CellProfiler (Jones et al. [2008]), CellCognition (Held et al. [2010])-, none of them allowed displaying selections of individual cellular positions and querying the atlas for co-expression domains with resolution at the cellular level.

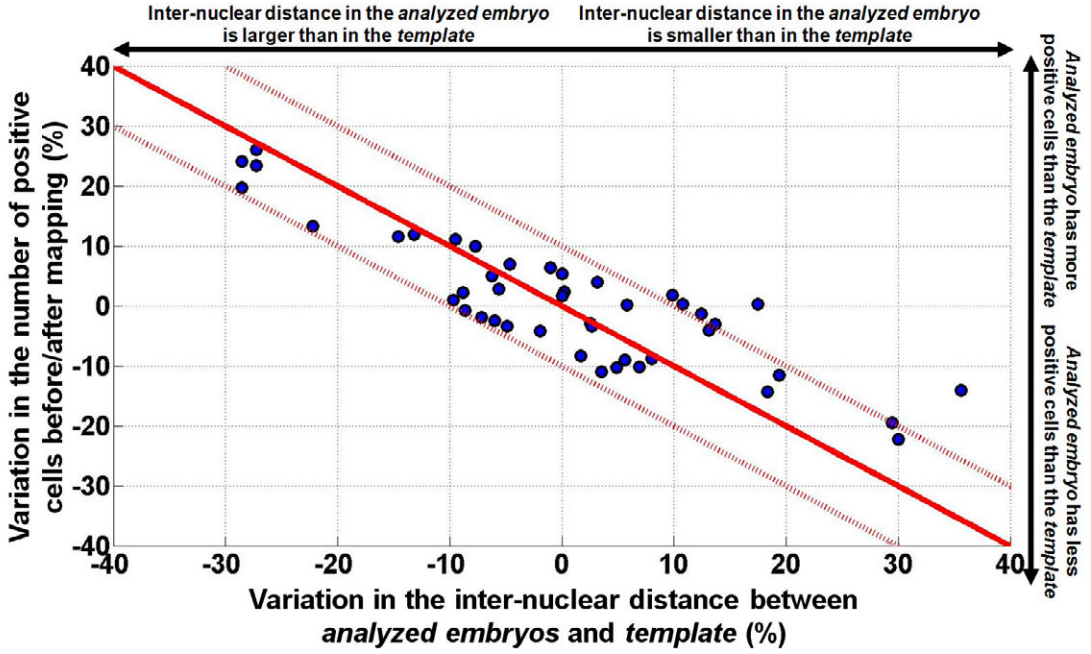

Figure 3.8: Variation in the number of *gsc* positive cells between *analyzed embryos* and *template* as a function of their relative internuclear distance. More than 95% of the *analyzed embryos* fall within a ± 10% deviation from the identity function, yielding a statistical p-value of 0.04. This experimentally establish the observed differences to reflect biological variability rather than methods inaccuracies. The two specimens deviating from this norm in the plot are 4.3 hpf. This suggests significantly higher variability in the number of *gsc* positive cells at the onset of gastrulation.

For this purpose, we developed and deliver the Atlas-IT interactive visualization interface (Fig. 3.9a) to explore the 3D atlas resource. It allowed us to interact with all the atlas data and superimpose raw data either as 3D volumes or orthoslices, segmented patterns and all the detected *template* nuclei or selected positive nuclei at each time point (Fig. 3.10). Atlas-IT, was used to assess the dynamics of genes co-expression domains or the variability of gene expression patterns (Fig. 3.11).

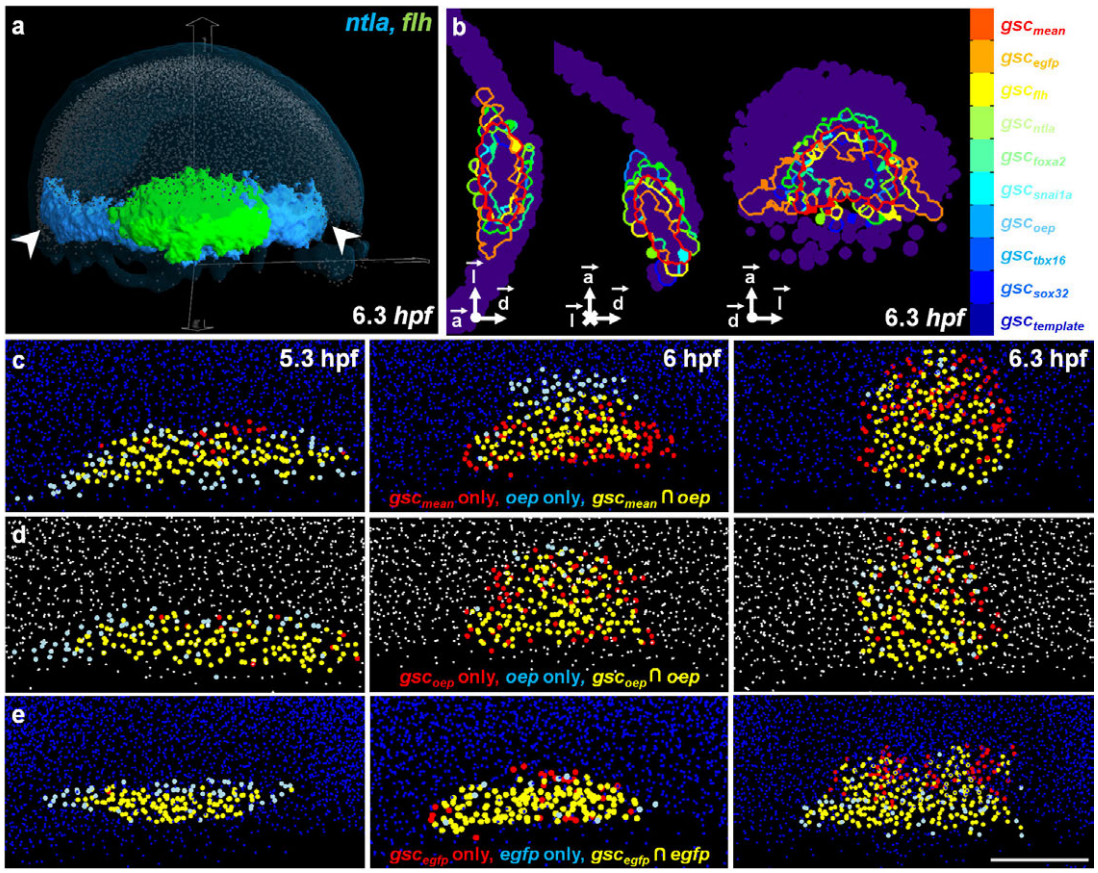

Figure 3.9: Exploring the 3D atlas with the visualization tool Atlas-IT. (**a**) Atlas-IT interface displaying the template nuclei in light blue, segmented gene expression patterns of *ntla* (in blue) and of *flh* (in green). (**b**) *From left to right*: equatorial, sagittal and dorsal views of the of the 9 individual *gsc* boundaries as compared to the mean *gsc* domain (in red) at 6.3 hpf. (**c**) Evolution of the gsc_{mean}-*oep* pair through time after being mapped onto the *template*. (**d**) Evolution of the gsc_{oep}-*oep* pair through time in the *analyzed embryo* where they were co-stained. (**e**) Evolution of the gsc_{egfp}-$egfp$ pair through time after being mapped onto the *template*.

3.5 Delivering a 3D atlas of gene expression in the early zebrafish embryo

We release a zebrafish early embryogenesis atlas comprising 9 gene expression patterns chosen to highlight gene expression dynamics underlying the formation of the Spemann organizer at the dorsal midline (Schier and Talbot [2005]): *goosecoid* (*gsc*), *casanova* (*sox32*), *spadetail* (*tbx16*), *one-eyed pinhead* (*oep*), *snail*

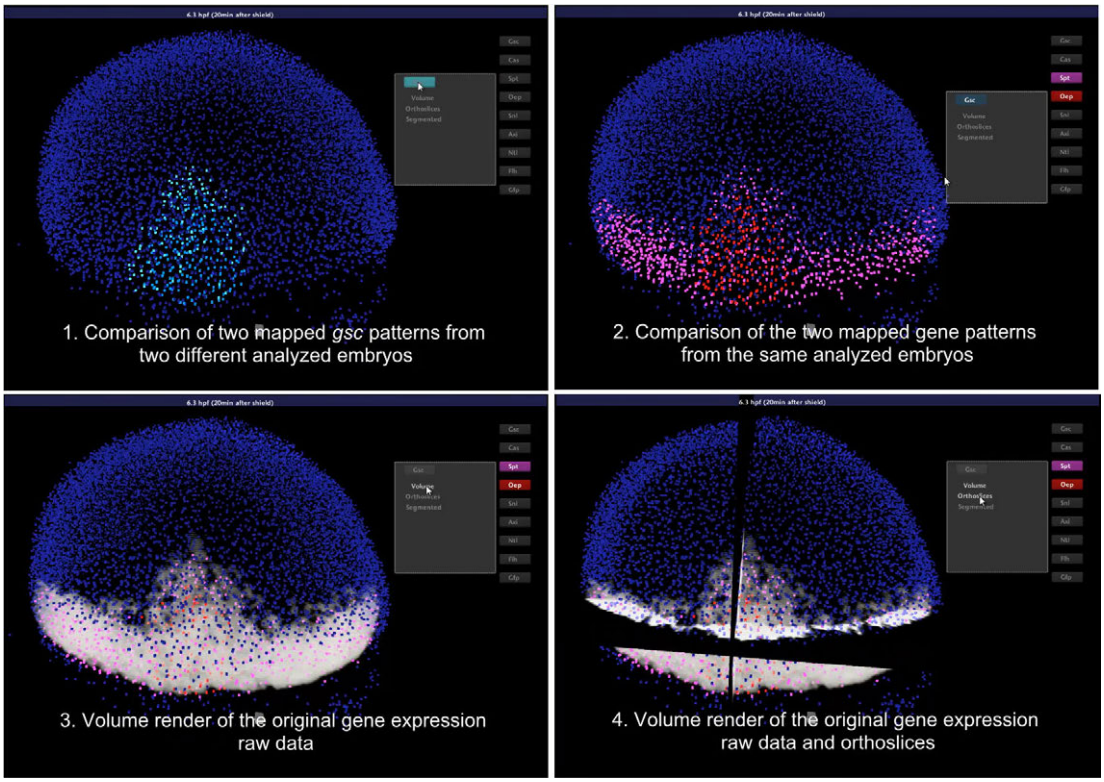

Figure 3.10: Visualization and supervision of the final atlas model with Atlas-IT.

(*snai1a*), *axial* (*foxa2*), *no tail* (*ntla*) and *floating head* (*flh*) in addition to *egfp* in a custom made transgenic line *Tg(-4gsc:egfp) isc3* (see section 3.2). The latter allowed us to validate this transgenic line as a consistent reporter of *gsc* gene expression (Fig. 3.9e) providing the means to further match the gene expression atlas and the 3D+time cell lineage (see chapter 5).

The atlas time series was chosen to explore gene expression dynamics from the onset of zygotic activation, i.e. 3 hpf until early gastrulation and encompasses the following time points: sphere (4 hpf), dome (4.3 hpf), 30% epiboly (4.7 hpf), 50% epiboly (5.3 hpf), shield (6 hpf) and late shield (6.3 hpf) according to the staging defined at 28.5 C (Kimmel et al. [1995]). For each new gene expression to be mapped, a cohort of 3 individuals was processed for double *in situ* hybridization and imaged. The methodology for the atlas construction was established by using one specimen of each cohort.

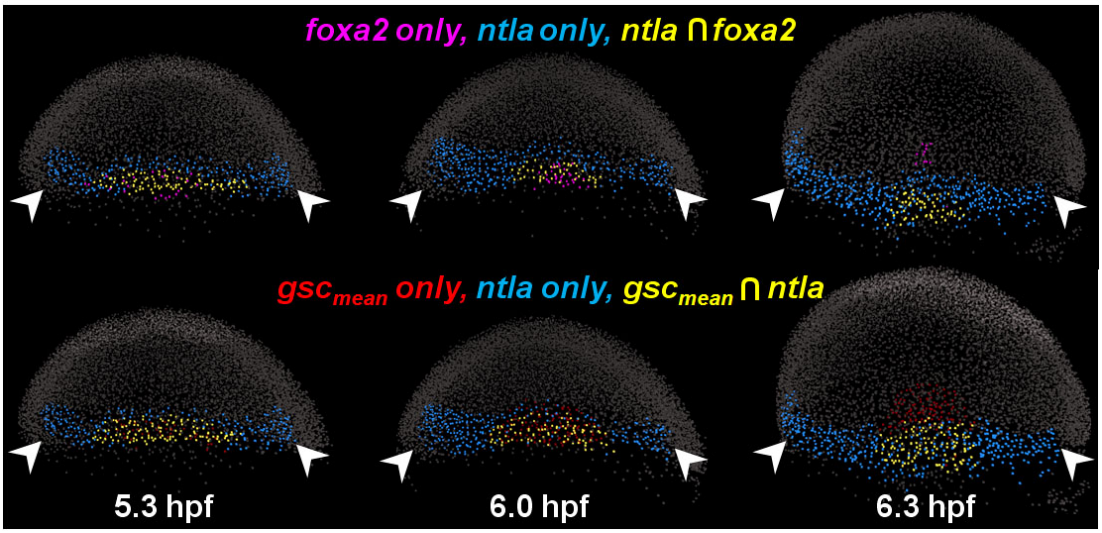

Figure 3.11: Evolution of the *foxa2-ntla* pair through time (top) and of the *ntla* gene expression domain against the mean *gsc* (bottom). Note that none of the *analyzed embryos* was co-stained for the *foxa2-ntla* pair making its comparison only possible in the atlas.

3.6 Evaluating the atlas accuracy and gene expression variance

In order to evaluate our approach, we confirmed that our atlasing strategy is robust against gene expression variability and preserves the spatial relationship between gene patterns whether assessed from co-staining in the same *analyzed embryo* or after mapping onto the *template*. *gsc* expression was revealed in 9 different specimens at each time step, i.e. 8 *analyzed embryos* and one *template*, providing a paradigmatic case to calculate a mean expression domain and assess gene variability.

To calculate a mean *gsc* expression, these 8 different *analyzed embryos* with *gsc* staining were mapped onto the *template*, where *gsc* expression had also been revealed for each developmental time point. Consequently, every nucleus, n_i, in the *template* can be assigned a value, V_i, ranging from 0 to 9, that indicates the number of different *gsc* patterns expressed by each cell (Fig. 3.12).

The difference between the surface of the gsc_{mean} expression and those of the individual *gsc* domains (gsc_j), Fig. 3.9b, was quantified with the mean Hausdorff

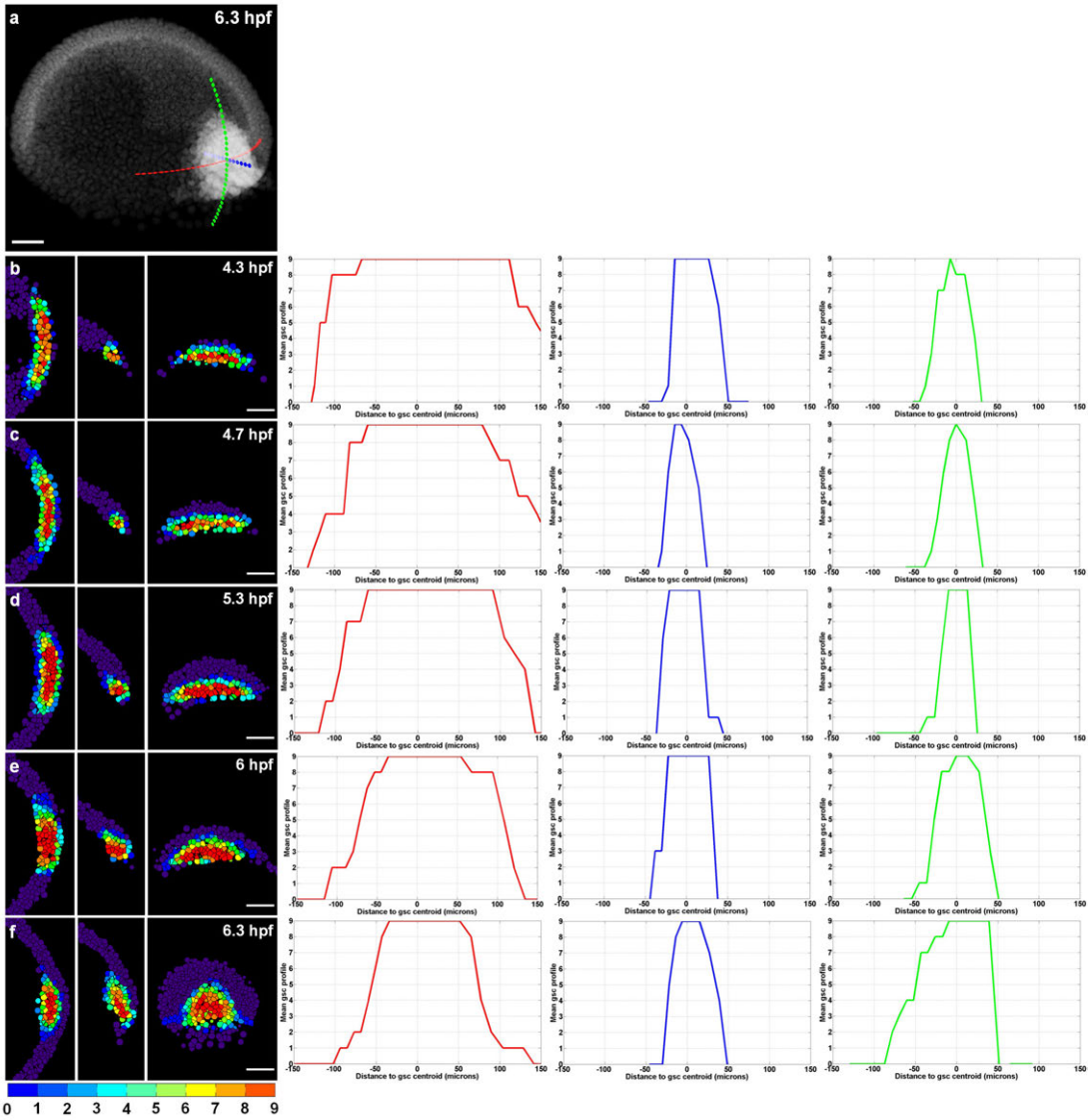

Figure 3.12: Inter-embryo variability of the *gsc* gene expression pattern. (**a**) Volume rendering of the mean *gsc* expression domain together with the three different orientations along which variability is measured: *lateral line* (in red), *radial line* (in blue) and *sagittal line* (in green). (**b-f**) *Left panel:* from left to right, equatorial, sagittal and dorsal orthoslices cutting the mean *gsc* expression domain at the level of its centroid. The color code indicates the number of individual *gsc* repetitions in the atlas cells based on the analysis of the 8 mapped *analyzed specimens* plus the *template*. *Right panel:* Profile showing how many embryos (out of the 9 specimens) expressed *gsc* in each cell along the three different orientations displayed in (a). Note the existence of a core domain (where all 9 cases expressed *gsc*) surrounded by successive layers of cells less frequently expressing this gene. We observe the variability of *gsc* expression, typically ranging between 1 and 4 cell rows, is not isotropic and correlates with the specimen's topological constraints and their time evolution, as shown by the profiles asymmetry. For instance, we can verify how *gsc* variability is always more prominent *laterally* than *radially*, where it is bounded by the embryo thickness. Similarly, the onset of *gsc* invagination at 6 hpf is responsible for an increasing asymmetry of the *sagittal* variability towards the animal pole.

distance. An average distance of less than 12 μm, corresponding to approximately one cell diameter, indicated both the accuracy of our mapping scheme and the inter individual variability observed for the shapes of the *gsc* expression domains (Fig. 3.13).

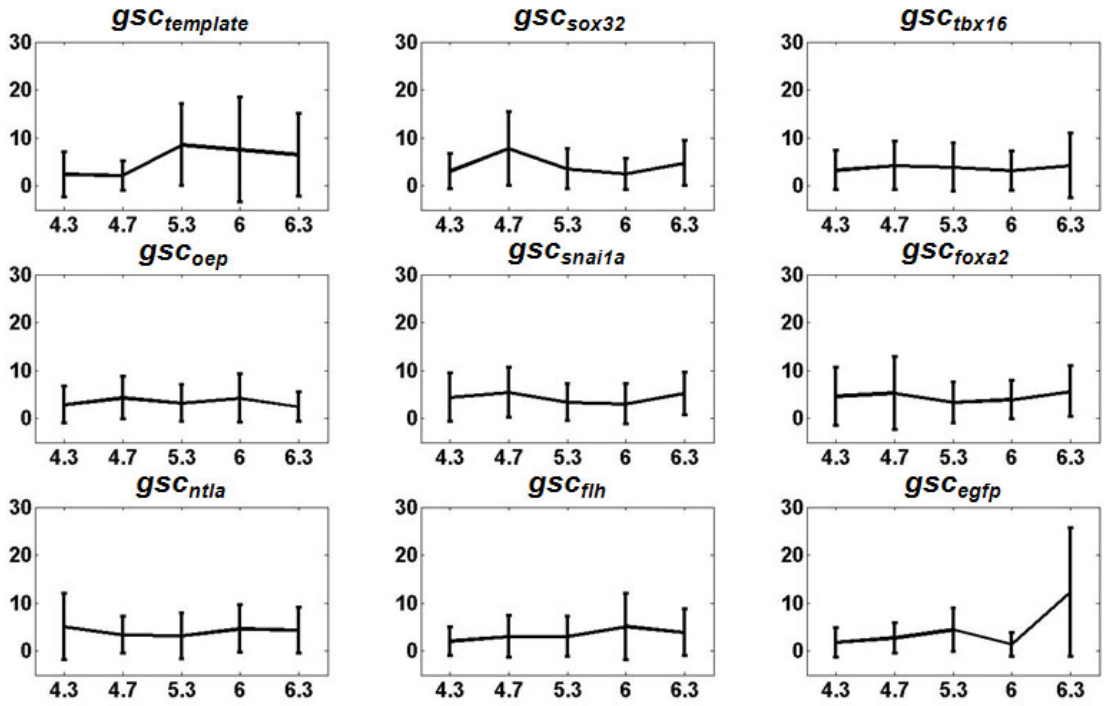

Figure 3.13: Quantification of patterns difference: application to the evaluation of *gsc* expression variance after mapping. Each of the individual *gsc* segmented domains mapped onto the atlas were compared to the mean *gsc* at every developmental stage between 4.3 and 6.3 *hpf*. The error bars represent the standard deviation of the Hausdorff distance along the borders. Numerical values are expressed in μm. The typical inter-nuclear distance ranges from 14 μm at 4.3 *hpf* down to 10 μm at 6.3 *hpf* (Fig. 3.6).

Additionally, we confirmed that, for all the gene pairs, the same spatial relationships were kept in the atlas with respect to the analyzed embryos (Fig. 3.14). That was the case, for instance, for the *oep*-*gsc* couple (Fig. 3.9d,e). There was an exceptional case, as the *gsc* pattern revealed in combination with $egfp$ (gsc_{egfp}) at 6.3 hpf extended laterally and symmetrically compared to the others and thus appeared as an outlier (Fig. 3.9b and Fig. 3.13). However, the difference in Dice's similarity coefficient for the pair gsc_{egfp}-$egfp$ in the *analyzed embryo* and

the pair gsc_{mean}-$egfp$ in the *template* still remained low (0.28). Apart from this outlier, the difference in Dice's coefficients for all the gene pairs were in the range of 0.08-0.16 validating the atlasing strategy and the information represented in the atlas.

Finally, Atlas-IT allowed us to visually verify that the gene domains location and relationships within the *analyzed embryos* are preserved after mapping onto the *template* (Fig. 3.15a,b,d,e). This virtual multiplexing is validated by the correspondence -within their biological variability- between the *gsc* patterns corresponding to different *analyzed embryos* (Fig. 3.15f). Consequently, our atlas provides the means to jointly compare gene expression domains that were not directly co-stained in any of the *analyzed embryos* (Fig. 3.15c).

3.7 Analytical methods and biological insights

The quantitative analysis of co-expression domains, their dynamics in space and time, the identification of synexpression groups (Niehrs and Pollet [1999]) and the clustering of cells according to their gene expression profile, are major scientific issues expected to be addressed by the atlasing strategy. To this aim, dedicated methods were further developed to achieve the quantitative analysis of genes co-expression patterns dynamics in space and time.

3.7.1 Virtual multiplex: Measuring co-expression between different gene couples

Overlapping gene co-expression (CC) was systematically analyzed for the 36 possible gene pairs through all the developmental stages (there are 9 genes and (9x8)/2=36 pairs). For every pair of genes (A and B), at each time point, we measured the number of cells that expressed both genes with respect to the total number of cells in each of the pair components: $CC_{A,B} = \frac{|A \bigcap B|}{|A|}$ and $CC_{B,A} = \frac{|A \bigcap B|}{|B|}$. This quantification was used for constructing a co-expression matrix (Fig. 3.16a) indicating precise aspects of patterns segregation which showed both expected features, such as a high co-expression between *gsc* and *oep*, and novel features, such as the temporally-robust similarity of the *gsc* and *egfp* expressions.

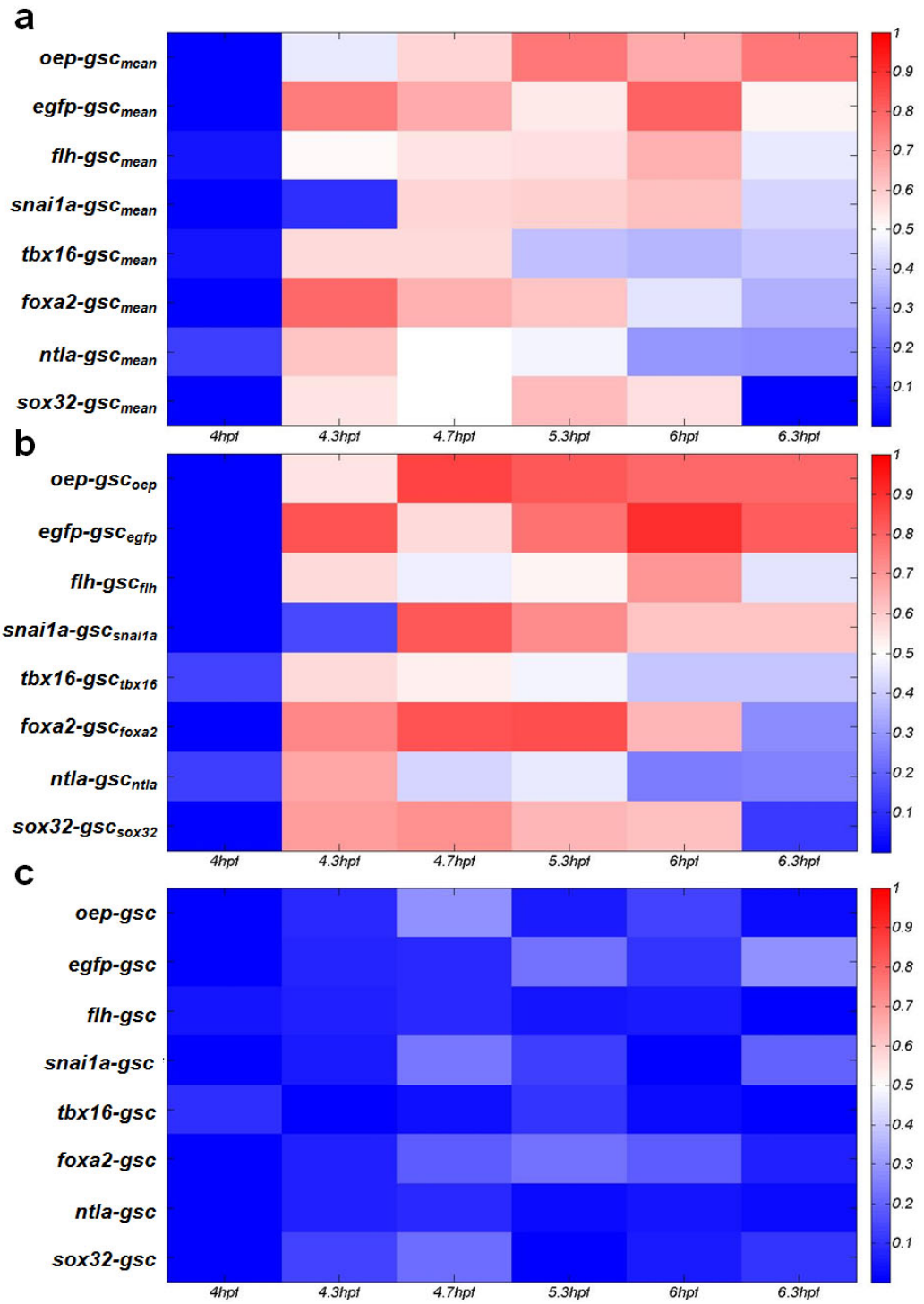

Figure 3.14: Comparison between the Dice's similarity coefficients in the *analyzed embryos* and the *template*. (**a**) Dice's similarity coefficients for the co-stained pairs mapped onto the atlas. (**b**) Dice's similarity coefficients for the co-stained pairs in the *analyzed embryos*. (**c**) Difference of the Dice's similarity coefficients between the co-stained pairs measured directly in the *analyzed embryos* and in the atlas. From this data, we conclude that the mapping procedure preserves the information provided by patterns co-staining in the *analyzed embryos*. The largest difference is found for the *egfp-gsc* pair at 6.3 hpf.

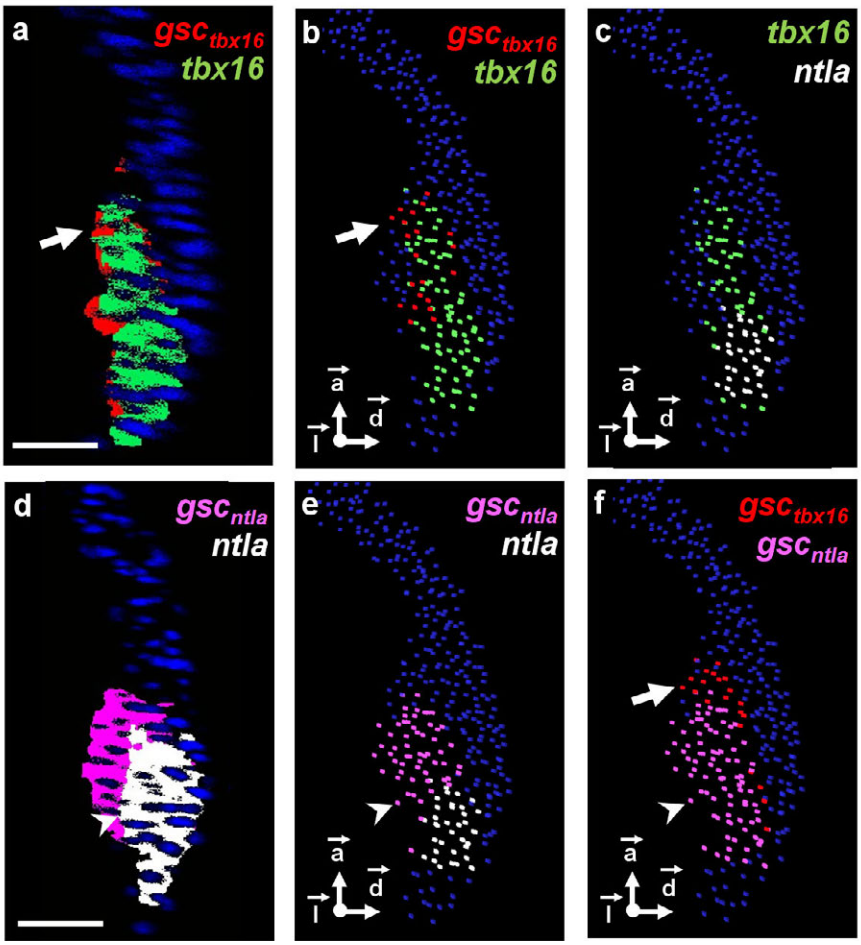

Figure 3.15: (**a, d**) Nuclei raw data (in blue) of two *analyzed embryos* together with the segmentation of their *gsc* (in red and purple respectively) and *tbx16* (in green) or *ntla* (in white) patterns. (**b, e**) Detected nuclei in the *template* (in blue) with positive cells for *tbx16* (in green), *ntla* (in white) and their accompanying *gsc* patterns (in red and purple, arrow and arrowhead respectively). (**c**) The *template* (in blue) displays positive nuclei for two different gene expressions, *tbx16* (in green) and *ntla* (in white), which were not directly co-stained in any of the *analyzed embryos* (digital multiplex). (**f**) The *template* (in blue) displays the positive nuclei for the two *gsc* patterns (in red and purple, arrow and arrowhead respectively) corresponding to the two different *analyzed embryos*. Scale bar, 100 μm.

Alternatively, this co-expression information was also employed to display the evolution through time of the topological relation between two gene patterns -i.e. identity, inclusion, exclusion, overlap- as a trajectory in a 2D space (Fig. 3.17). As an example, the trajectory of genes *oep* and *sox32* goes from inclusion at 4.3

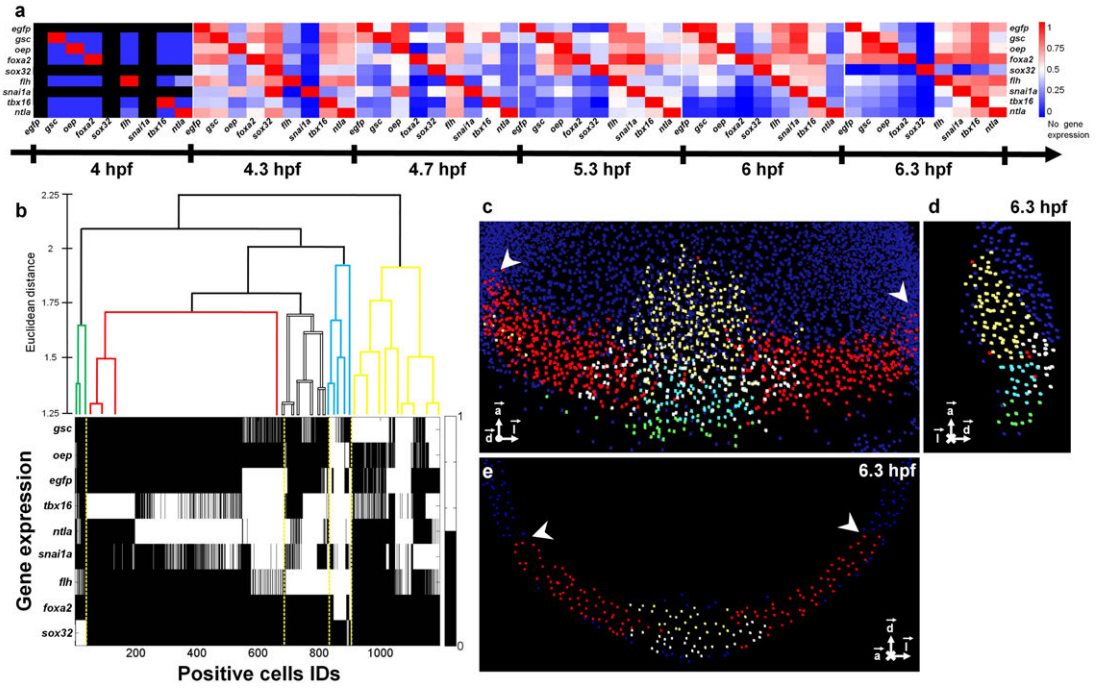

Figure 3.16: Assessing genes co-expression and cells genetic profile. (**a**) Matrix displaying the percentage of cells co-expressing the genes of a given pair at stages from 4 to 6.3 hpf (see also Fig. 3.17). Gene pairs were ordered according to the similarity of their patterns evolution in time (see Fig. 3.19). (**b**) Clustering of 1,194 positive cells according to the similarity of their gene expression profile at 6.3 hpf. (**c**) Volume rendering, (**d**) lateral view and (**e**) coronal view of *template* cells at 6.3 hpf. Cells were classified according to their gene expression profile which revealed 5 distinct morphogenetic domains: dorsal hypoblast (yellow), marginal dorsal epiblast (blue), dorsal epiblast (white), paraxial and lateral blastoderm margin (red), forerunners and dorsal YSL (green). White arrowheads indicate the limits of the imaged analyzed embryos. Scale bar, 100 m.

hpf, to intersection between 4.7 and 6 hpf to finally exclude each other at 6.3 hpf (Fig. 3.17b).

Likewise, we also measured the Dice's coefficient (D) as an indicator of co-expression similarity for the 36 possible pairs through time: $D_{A,B} = \frac{2|A \bigcap B|}{|A|+|B|} = D_{B,A}$. The Dice's coefficient is symmetric with respect to the pair of genes that are measured ($D_{A,B} = D_{B,A}$), allowing a more compact representation of the gene expression evolution (Fig. 3.18) than the measurement obtained by the co-expression matrix (Fig. 3.16a), where $CC_{A,B} \neq CC_{B,A}$. Dice's representation

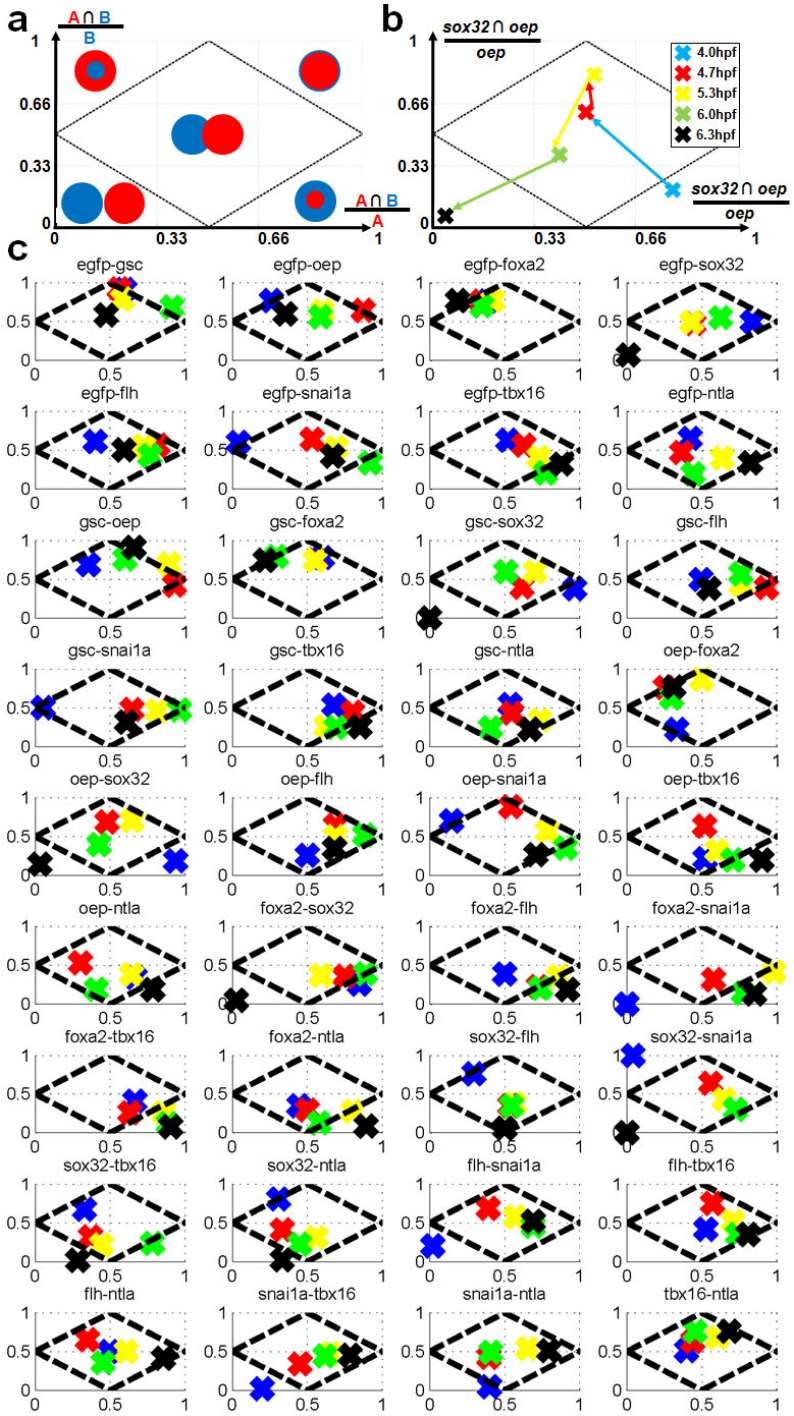

Figure 3.17: A synthetic view of gene pairs co-expression and its evolution through time. (**a**) Gene pairs expression can fall into 5 possible categories defined by gene patterns similarity: A and B expression domains exclude each other (bottom left), A is included in B (bottom right), B is included in A (top left), A is identical to B (top right), A and B domains partially overlap (center). (**b**) This chart allows to visualize the segregation of the *oep-sox32* expression through time. (**c**) Gene patterns relationships and their evolution in time for the 36 possible pairs. Coherence with a priori knowledge was checked and demonstrates the power of the atlas construction strategy and further analysis tools.

highlighted the similarity of *gsc* and *egfp* patterns until early gastrulation. This feature was captured by the high values of the Dice's coefficient. The *gsc-egfp* pair analyzed achieved a mean Dice's coefficient of 0.77 through time with a standard deviation of 0.1, validating the transgenic line as an acceptable reporter of the *gsc* activity at early developmental stages.

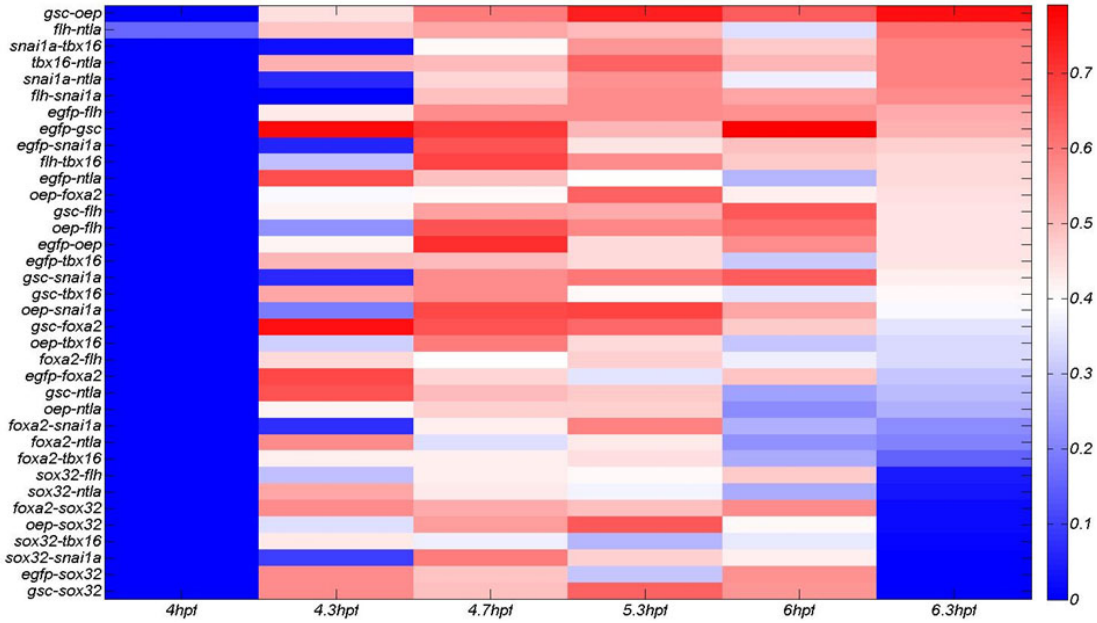

Figure 3.18: Evolution of the Dice's similarity coefficient for all the possible gene pairs. The Dice's similarity coefficient: $D_{A,B} = \frac{2|A \bigcap B|}{|A|+|B|}$ was calculated for the 36 gene pairs. They are arranged in descending order, from top to bottom, according to their Dice's similarity coefficient at 6.3 hpf.

3.7.2 Clustering cells according to their gene expression profile

To study the relationship between the morphological positions of cells within the embryo and their gene expression profiles for a given developmental stage, *template* cells were grouped according to the similarity of their gene expression profiles using a hierarchical clustering scheme (Eisen et al. [1998]). Although this clustering does not use any a priori information about the cellular spatial location, the resulting categorization according to each cell expression profile

defined distinct morphogenetic domains for each developmental stage.

In particular, for each of the M *template* cells, c_i, observed at one certain time, t, we associate a gene expression vector:

$$\overrightarrow{c}_i = (g_{i,1}, ..., g_{i,N})$$

where N=9 is the number of genes in our case and $g_{i,j}$ is set either to 1, if the i-th cell expresses the j-th gene, or to 0 otherwise. For instance, a cell with id i=7 that expresses only two genes, *gsc* (j=1) and *snai1a* (j=4) will be characterized by the vector $c_7 = (1, 0, 0, 1, 0, 0, 0, 0, 0)$.

A weighted pair group method using arithmetic averages (WPGMA) was then performed on these vectors using the Euclidean distance. This method was implemented using the Statistics toolbox in Matlab (The Mathworks, Inc., USA).

The analysis was performed separately for each time of observation and is best illustrated at 6.3 hpf (Fig. 3.16b-e) with the classification of 1,194 cells according to their gene expression profile $\overrightarrow{c}_i = (g_{i,1}, ..., g_{i,9})$.

This resulted in the clustering of cells into 5 spatial domains with characteristic gene expression profiles: dorsal hypoblast (yellow), marginal dorsal epiblast (blue), dorsal epiblast(white), paraxial and lateral blastoderm margin (red), forerunners and dorsal YSL (green), see Fig. 3.16c-e. This clustering strategy revealed the antero posterior and medio lateral patterning of the mesendodermal tissue at the onset of gastrulation.

The proposed cellular clustering constitutes a tool that permits to check whether expression profiles are linked to specific morphological locations. In other words, it opens up the possibility to explore the link between phenotype and genotype at the cellular level.

3.7.3 Clustering genes according to their spatio-temporal expression patterns

We introduce a clustering of genes according to the spatio-temporal characteristics of their expression pattern with the aim of refining the identification of synexpression groups, i.e. genes with potentially the same spatial and temporal

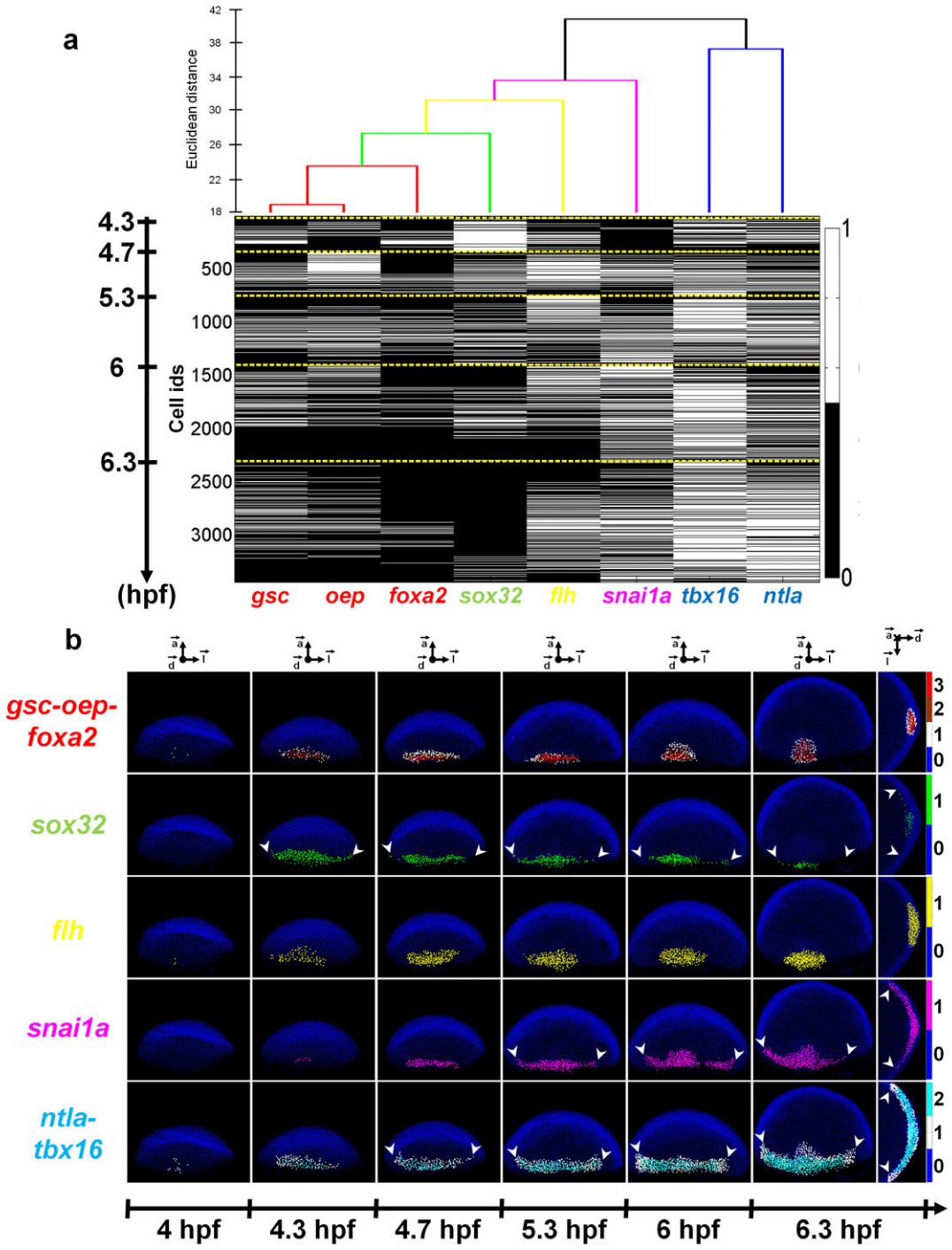

Figure 3.19: Genes synexpression groups defined by their spatio-temporal clustering patterns. (**a**) Genes hierarchical clustering according to the similarity of their spatio-temporal regions of expression defined 5 different groups with characteristic spatio-temporal behaviors. For each group, a color code (column right of the panel) was displayed to indicate whether cells expressed 0, 1, 2 or 3 genes. The 8 genes analyzed fall into the following synexpression groups: *gsc-oep-foxa2*, *sox32*, *flh*, *snail1a*, *ntla-tbx16*. (**b**) Visualization of the synexpression groups identified in (a). Arrowheads indicate the limits of the imaged volume in the *analyzed embryos*.

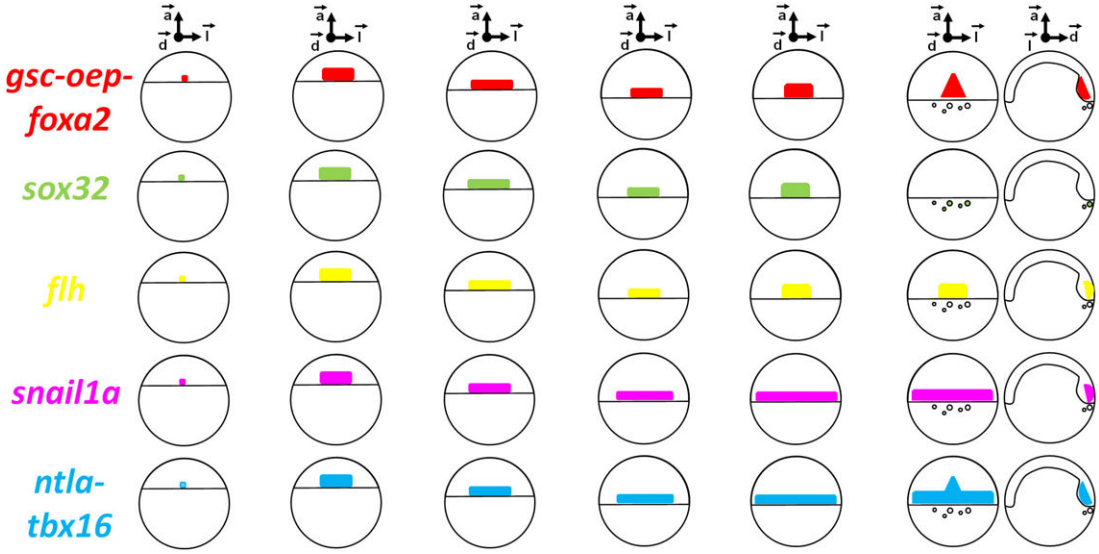

Figure 3.20: Simplified dynamics of each of the 5 spatio-temporal gene expression clusters displayed in Fig. 3.19.

regulation of expression. This spatio-temporal clustering of gene expression allows to automatically group together gene patterns according to the spatiotemporal similarity of their 3D+time expression pattern at the single cell level.

While the cellular spatial clustering (section 3.7.2) grouped cells according to their gene expression profile at a given developmental stage, the clustering proposed in this section, groups together genes that express at similar cellular locations across all developmental stages. For each of the N=8 genes under study, say the i-th gene, we associated a spatiotemporal region of expression vector:

$$\overrightarrow{r}_i = (c_{i,1}, ..., c_{i,P})$$

where P is the total number of cells observed across all observation times. Here:

$$P = \sum_{k=1}^{6} M_k = 55,759$$

where M_k is the number of cells observed at the k-th time of acquisition. For the i-th gene, the j-th coordinate of that vector, is set to 1 if the j-th cell expresses

that i-th gene, and set to 0 otherwise. For instance, a gene "*test*" which is just expressed in a couple of cells (with ids j=1,2) at the first time point and in one cell (with id j=55,759) at the last time point will be characterized by the following spatiotemporal region of expression vector: $r_{test} = (1, 1, 0, 0, ..., 0, 0, 0, 1)$.

Clustering the genes according to their associated region of expression vectors is performed with the same WPGMA algorithm as before, again using the Euclidean distance and Matlab implementation.

8 genes in the atlas (*gsc*, *oep*, $foxa2$, $sox32$, flh, $snail1a$, $tbx16$, $ntla$) were clustered into 5 classes that defined 5 synexpression groups compatible with biological descriptions described in previous literature (Kudoh et al. [2001]), see Fig. 3.19 and Fig. 3.20.

3.7.4 Shannon's entropy of gene expression

We propose to use Shannon's entropy (Cover and Thomas [2006]) to measure cells' gene expression profile complexity and its temporal evolution. We define the genetic entropy as the average number of bits needed to code the gene expression profile of a cell at a given developmental stage.

We consider that the gene expression observed at the i-th cell ($1 \leq i \leq M$, where M is the number of cells) is the value taken by a discrete random variable G_i, the possible values for G_i being all the 2^N N-uples $(g_{i,1}, ..., g_{i,N})$ of 0s and 1s, where N is the number of gene products considered in the atlas (N=9 in our case, and reading each N-uple as a binary number, we identify the set of all possible N-uples with the set of integers from 0 to 2^{N-1}), the value of the j-th digit being set to 1 if the j-th gene is expressed by that cell and set to 0 otherwise. Assuming that the random variables $G_1, ..., G_M$ are independent and identically distributed, say with the same law as a random variable G, their entropy is:

$$H(G) = -\sum_{0 \leq k \leq 2^{N-1}} p_k \log_2 p_k$$

with the usual conventions: p_k denotes the probability of the event $G = k$ and we set $0.log_2 0 = 0$, where 2 was chosen as the base in order to express the result in bits. Each p_k can be estimated from the observed sample $(g_1, ..., g_M) =$

$((g_{1,1}, ..., g_{1,N}), ..., (g_{M,1}, ..., g_{M,N}))$ by:

$$\hat{p}_k = \frac{n_k}{M}$$

where n_k is the number of *template* cells showing the k-th expression N-uple and M is the total number of cells in the *template*. Replacing each p_k in the formula by the corresponding $\hat{p}_k$ gives an estimate for $H(G)$. Under the same hypotheses the total entropy for the population of M cells is equal to $M.H(G)$. Note that if the G_i are not independent, the total entropy of the population (defined using a single random variable to generate the combined expressions of all the cells in a population, and requiring the observation of several populations to permit estimation) can only be less than or equal to $M.H(G)$.

In our atlas, each cell can express 1 of the $2^9 = 512$ possible gene expression profiles. We measured the entropy of our atlas and observed a rapid increase to a peak of 5.7 bits at 4.7 hpf, then a slight decrease until 6.3 hpf (Fig. 3.21a). The increase of entropy with time could be the progressive increase in the number of expressed genes (Fig. 3.21b). In addition, we measured the contribution of each gene expression profile to the global entropy. We observed that, per time step, only around 100 different gene profiles were expressed out of the possible 512. Moreover, we found that during the period analyzed, from 30 to 50 gene profiles were responsible for 75% of the whole entropy (Fig. 3.21c,d).

The concept of genetic entropy can be useful to identify changes in gene expression, e.g. a new gene expression, independent from the other expressions, that would suddenly start expressing at a certain time step would imply an increase of information of 1 bit. Similarly, we could think of a mutant embryo, having the exact genetic code as his wild type counterpart except for one inhibited gene expression. Assuming this gene is independent from the other expressions in the wild type, the entropy of the gene profile in the mutant would be one bit lower. Gene expression entropy as an information measure can also be used as an indirect measure of redundancy among the studied gene expressions. In our case, supposing an extreme case where all the possible 2^9 gene expressions profiles were evenly distributed along the cells, the gene expression entropy would be 9 bits. However, the measurements are below 6 bits of information, meaning that

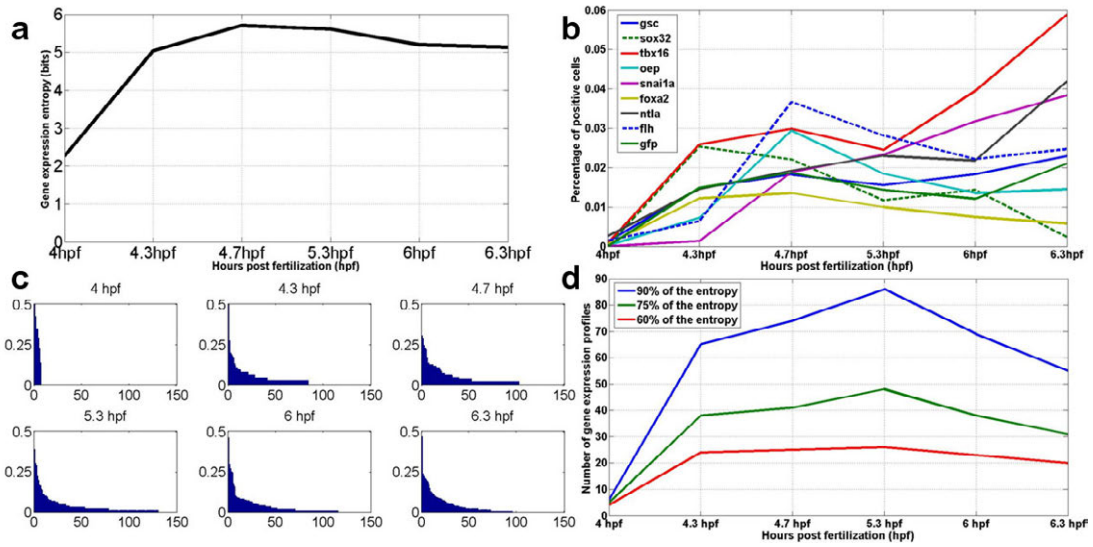

Figure 3.21: Gene expression entropy. (**a**) Gene expression entropy as a function of time: Shannon's entropy provides a measurement of cells' gene expression profile complexity. (**b**) Percentage of positive cells for each gene expression as a function of time. A gene expression, inhibited until a certain time step, that would suddenly start expressing would, for instance, make the entropy increase 1 bit. (**c**) Information, in bits, provided by each gene expression profile at each time step. We can observe that many of the possible 2^9=512 possible gene expression profiles are, in fact, never used and that most of the information is conveyed by a small number (from 30 to 50) of representative combinations. (**d**) Number of gene expression profiles required to convey 60% (red line), 75% (green line) and 90% (blue line) of the total entropy at each time step.

-by just taking into account our small universe of 9 genes- there is a 3 bits redundancy. Theoretically, from the point of view of information theory, we could have obtained the same gene expression biodiversity by using a synthetic gene expression profile with just 6 genes.

3.8 Discussion

We developed and now deliver the software packages Match-IT and Atlas-IT dedicated to the reconstruction, analysis and visualization of a 3D atlas of gene expression in the early zebrafish embryogenesis. The atlas comprises 6 different time points between 4 and 6.3 hpf and gathers data for 9 gene patterns into 6 different *templates*.

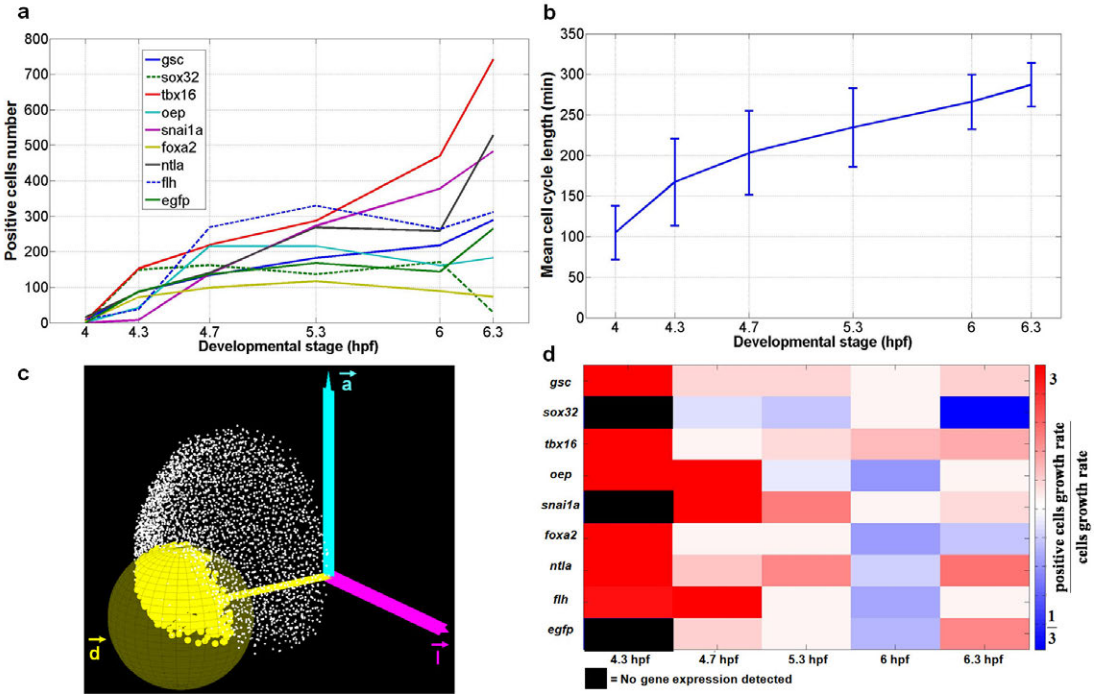

Figure 3.22: Correlation between the cell proliferation and the gene expression domains expansion. (**a**) Evolution of the number of positive cells for each of the 9 considered gene products. (**b**) Temporal evolution of the cell number in the region of interest (ROI) centered on the dorsal side of every *analyzed embryo* as shown in (c). The cell proliferation rate extracted from this experiment matched previous observations (Keller et al. [2008]). (**c**) Dorsal region of interest (ROI) used to measure the cell proliferation rate in every *analyzed embryo*. (**d**) Ratio between the increase rate of the positive cells for a given gene and the estimated overall cell proliferation rate. This ratio indicates whether the gene expression pattern dynamics can be explained by sustained expression in proliferating cells or requires up regulation (such as for *egfp* by 6.3 *hpf*) or down regulation (such as for *sox32* by 6.3 *hpf*).

So far, the only method delivered for the reconstruction of gene expression atlases in the zebrafish was designed for the brain at late developmental stages where morphological landmarks can be found (Ronneberger et al. [2012]). Given the complexity of building a zebrafish brain atlas at late stages, Ronneberger et al. [2012] imposed strong constraints on the data in terms of staining protocols and imaging (Rath et al. [2012]). Our atlasing strategy was made to map partial 3D volumes onto whole embryos chosen as *templates*. Specimens are only required to display in addition to any pattern of interest, nuclear staining for

single cell counterstain and a common gene expression, *gsc* in the present version of the atlas, serving for the registration step. This gene was chosen as a relevant marker, with early, strong and well-regionalized expression, to serve as a reference for constructing the dorsal side gastrulation atlas. We thus have minimal prerequisite for data format and specimen preparation (described in section3.2) and this should facilitate the introduction of new data into the atlas. In addition, our scheme would be easily applied to other model organisms at early steps of development when too few morphological landmarks are available to use landmark-based registration methods for mapping analyzed patterns onto the atlas *template*. The possibility of visual inspection and, if necessary, manual correction using the Match-IT graphical interface contributes to flexibility and accuracy when integrating new data into the atlas and validating the result.

Our choice to work at the cellular level with an automated and supervised method of nuclear center detection proved being suitable to quantify features of gene expression patterns dynamics at the single cell level. This opens up the possibility to study whether cell proliferation alone is enough to account for the gene expression pattern expansion, which is particularly relevant for early zebrafish development when cell divisions happens at constant global cell volume, thus correlating internuclear distance and cell proliferation rate (Fig. 3.22). On the other hand, the cellular resolution of the atlas is required to exploit the correlation between gene expression dynamics and the cell lineage. Indeed, the single cell resolution would allow to further map the atlas onto digital specimens reconstructed from live in toto imaging starting with our *Tg(-4gsc:egfp) isc3* transgenic line, see chapter5.

Working at the single cell resolution was also meant to tackle the problem of gene expression quantification. Current strategies used for *in situ* hybridization could at best provide relative measurements within each analyzed specimen, which are suitable for quantifying graded patterns and fuzzy borders. Such a relative quantification would be readily available from our atlas. We expect future developments of the programmable *in situ* amplification technique (Choi et al. [2010], Mahou et al. [2012]) to help achieving the absolute quantification of gene expression at the single cell level making it comparable among different analyzed specimens.

The relevance of the atlas relies on its ability to represent and integrate the same information as would be obtained by inspecting different patterns in the same specimen. This depends on the accuracy of the registration strategies but most importantly on how the atlas construction scheme deals with individual variation. Every step of the mapping strategy has to cope with the individual variation in terms of shape, cell number, cell density, variability of the reference gene pattern. In this context, the choice of the *template* is crucial. The *templates* should be chosen to be the closest to the mean of the population, based on geometric parameters and gene expression. Ideally, a multiscale model of individual variation should drive the choice of the atlas *templates* as well as representative references to guide the mapping. In our case the *gsc* pattern served to guide the registration step, based on the hypothesis that its expression is symmetric to the bilateral plane. Although this is a reasonable assumption, it is an approximation that might be confronted to other features such as more reference gene patterns or additional morphological features. In this context, we calculated a mean *gsc* expression after registering the domains from 9 different specimens. This mean *gsc* pattern could be subsequently used as a new reference to refine the global mappings. Moreover, all the genes gathered in the atlas could be averaged, thus preventing potentially misleading conclusions based on single specimens that might be outliers. The increase in size of the cohorts will allow to explore the possible convergence of the averaging strategy toward a single (or multiple) prototypical specimen.

The atlas resources will only be fully exploited with the development and use of automated analysis methods and dedicated visualization tools. We developed Atlas-IT to provide a number of functionalities, not available in any of the visualization tools that we explored, to augment/visualize/analyze raw data, segmented data, mean gene expression domains, genes co-expression patterns, synexpression groups and morphogenetic domains revealed by cells clustering. Interactive visualization and data display are essential to reveal biologically relevant information. Exploring analytical methods to highlight spatial and temporal correlation is also a major issue. Clustering methods are typically used to establish cells and tissues gene expression profiles from microarray data (Eisen et al. [1998]). More recently, clustering strategies have also been used to group anatomical regions according

to their gene expression profile (Lein et al. [2007], Liu et al. [2009], Frise et al. [2010]). In this paper, clustering methods based on gene expression profiles at the single cell level, $(x, y, z, t, g_1, ...g_N)$, are applied for the first time and provide means to reveal morphogenetic domains and synexpression groups. Additionally, the concept of genetic entropy applied to the analysis of atlases data introduces a new systematic way to assess cells diversification and its underlying genetic complexity.

Future expansions include substituting double *in-situ* acquisitions by multiplexed *in-situs* (Choi et al. [2010]) and/or multi-color imaging (Mahou et al. [2012]). This would open up the possibility of including further constant reference patterns that could increase the mapping accuracy and robustness. In addition, the atlas construction process would also benefit from including multiple gene products in one single mapping process. The proposed strategy is currently being tested in different embryo strains, including *zoep* mutants, opening up the possibility to run wild type-mutant comparisons within the same model. Future work also includes extending the current atlas to later developmental stages up to *bud* (10 hpf) and the development of robust gene quantification schemes (Crombach et al. [2012a], Dubuis et al. [2013]) that allow to overcome the current "on-off" segmentation model by extracting standard gene expression quantifications that are robust within specimens coming from different batches imaged under different acquisition conditions.

Developing atlases is a necessary step to integrate multiscale and multimodal data organized, displayed and annotated to provide and share as much relevant information as possible. Developmental biology remains far behind the biomedical field for constructing and mutualizing such resources. This means that before reaching a consensus and establishing standards in the discipline, a broad field remains to be explored in terms of different schemes, their flexibility, their potential and limitations. The atlas construction process itself leads to tackling some of the most difficult biological questions linked to the individual variability, its components and characteristic scales. On the other hand, methods and resources will spread and grow only if deployed together with adequate querying and analytical tools. The methods and the first release of the zebrafish blastula and early gastrula atlases proposed here are meant to contribute to the reconstruc-

tion of the zebrafish embryonic physiome in different genetic and environmental conditions.

This uneasiness comes over me from time to time, and I feel as if I've somehow been pieced together from two different puzzles.

Haruki Murakami

Chapter 4

A framework to reconstruct a 3D atlas of gene expression in the late zebrafish brain

4.1 Introduction

In this chapter, the main objective is to develop computational methods to reconstruct the cartography of gene expression in the zebrafish larval forebrain acquired from *in situ* hybridization microscopy images. This cartography will serve to relate the morphology and functionality of the different forebrain areas with their corresponding gene expression profiles. In other words, the aim revolves around providing the computational tools to answer the following question: How is neuronal identity related to its spatio-temporal history in terms of combined gene expression patterns?

So far, comparison of gene expression required pair-wise *in situ* hybridization in individual experiments. In this chapter, we introduce a software pipeline that automatically maps gene expression data with cellular resolution to a standard larval zebrafish brain.

The resulting resource, made of cell populations defined by their genetic profile, will help understand normal brain development and neuronal differentiation as well as phenotypes affecting nervous system formation and function. This

will be achieved by mapping gene expression domains at high resolution to neuroanatomical structures that will enable us to use digital 3D gene expression information to correlate with neural phenotypes and functional domains.

In previous related works, Peng et al. [2011] developed a registration approach based on landmarks and thin-plate splines to align fly brains but this approach has not been tested on data from vertebrate animal models. Ullmann et al. [2010] developed a standard 3D atlas of the zebrafish brain anatomy from a MRI acquisition of a single specimen which did not include gene expression data. Recently, Ronneberger et al. [2012] introduced a framework for the automatic detection of zebrafish brain landmarks and their subsequent alignment using a landmark-based registration refined by an elastic intensity-based method. This framework could not be adapted to our dataset specifications as it relies on a sophisticated image acquisition protocol (Rath et al. [2012]). In particular, this protocol implies taking 16 different images per acquisition at two different wavelengths, two laser intensities, at two positions (front and rear) from two sides (dorsal and ventral). The design principle of our approach is to avoid such requirements employing an intensity-based diffeomorphic registration scheme which does not require complex acquisition schemes or any specific labeling, relying only on the nuclei channel to guide the gene expression registration. Both the use of diffeomorphisms and the introduction of gene expression data have been recently identified as two of the major challenges in atlases construction (Evans et al. [2012]). Compared to our work on early zebrafish development (see chapter 3), the processing of later developmental stages implies facing two major technical issues: On one hand, the apparition of distinct morphological landmarks leads to the use of non-rigid transformation strategies that can accurately align these distinct forebrain structures of the *template* and the cohort of individuals. On the other hand, the size of the late zebrafish brain leads to signal depth-extinction issues that were tackled by acquiring two different views, ventral and dorsal, which must be registered and fused together.

Finally, we aim at gathering expression patterns with resolution at the cellular level for three different developmental stages: 24, 30 and 48 hpf. The resulting atlas will be adapted for visualization with Atlas-IT and a regionalization of the *template* will be made available to allow queries for gene co-expressions.

4.2 Image acquisition

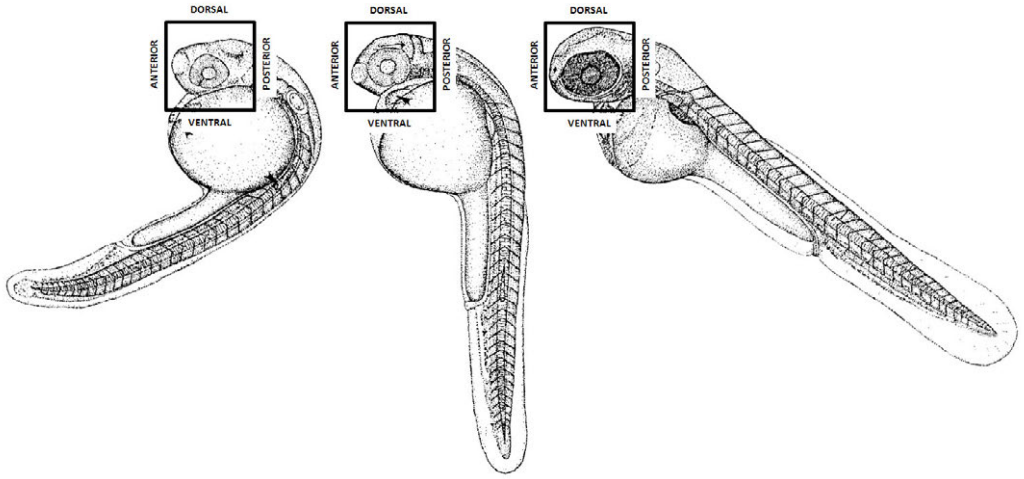

Figure 4.1: Zebrafish embryo development at 24, 30 and 48 hpf (Kimmel et al. [1995]). The squares have an approximate length of 300 μm and indicate the imaged area.

Image acquisition was performed by a Zeiss confocal microscope LSM 700, 40X/1.3 NA oil objective with a spatial resolution of 0.31x0.31x1 μm.

Four-color labeling was performed on paraformaldehyde-fixed zebrafish embryos which resulted in four-channel images per acquisition: nuclei (*DAPI* staining), a first gene expression pattern, a second gene expression pattern and the *th* (Tyrosine hydroxylase) marker (see Fig. 4.2). The *DAPI* staining was performed for every acquired dataset and is the only requirement to guide the registration process (see section 4.3). The three other fluorescence channels were used to detect mRNA expression of three given genes by *in situ* hybridization. These patterning genes were systematically associated to *Th* (Tyrosine hydroxylase) a marker of dopaminergic nuclei.

Three different developmental stages were analyzed: 24, 30 and 48 hours (Fig. 4.1). Two different views, ventral and dorsal (Fig. 4.2), were imaged to avoid depth-signal extinction issues and cover all the depth of the zebrafish brain. For each imaged individual, only one of these views was acquired, meaning that none of the ventral/dorsal acquisitions belong to the same specimen except in the case of the selected 3 *templates*, one for each developmental stage, for which both their

ventral and dorsal views correspond to the same specimen.

Finally, 8 different gene expression patterns, characteristic for the forebrain regionalization, were chosen: *nkx2.1a*, *pax6a*, *dlx2a*, *sim1*, *tbr1*, *rx3*, *six3b* and *dbx1a*.

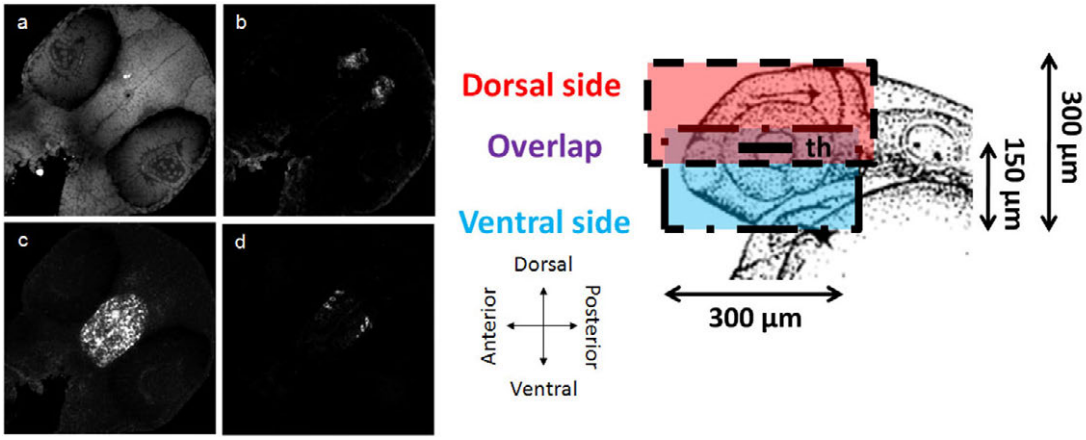

Figure 4.2: *Left panel*, Typical confocal imaging example of a fluorescent *in situ* hybridization (ISH) on a 48 hpf zebrafish brain. A single plane (1 μm) of a ventral view is shown here. *DAPI* staining allows detection of individual cells as well as anatomical landmarks (**a**). The three remaining channels were used to detect two gene expression patterns: the *sim1* expression in the preoptic area (**b**) and the *nkx2.1a* expression in the hypothalamus (**c**) and the *th*-expressing neurons in the posterior tuberculum (**d**). *Right panel*, dorsal and ventral views, with a certain overlap between them, are acquired for different zebrafish brains. This acquisition set up allows obtaining a complete map of the whole brain while avoiding the typical optical microscopy problems with signal depth extinction.

4.3 Methods

We divide the brain atlas reconstruction strategy in two major steps: First, we employ an initial alignment (based on manual landmarks) subsequently followed by an affine and a non-rigid registration schemes in order to gather datasets coming from the same acquisition view into their corresponding ventral (or dorsal) *template*. Then, given the fact that both the ventral and dorsal views of the *template* belong to the same specimen, we rigidly register and fuse both sides to create the complete zebrafish brain atlas (Fig. 4.4). We describe these procedures

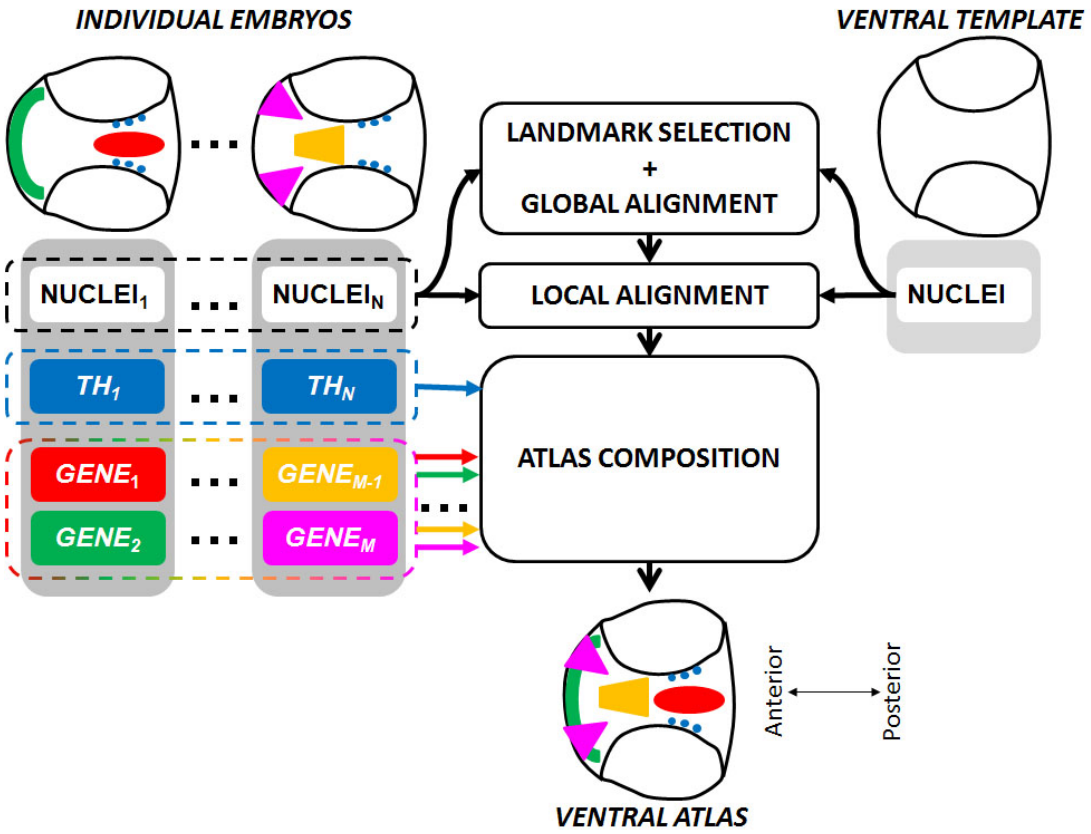

Figure 4.3: Schematic illustration of the brain alignment workflow to build an atlas of the ventral side.

in detail in the following sections (Fig. 4.3).

4.3.1 Global alignment

At first, we tested a tentative global alignment scheme based on the labeling of the *th* neurons, which are symmetrically placed at both sides of the anterior-posterior axis. However, given that the *th* markers position was not completely stereotypical, this procedure did not result to be a reliable guide to perform this initialization step.

Consequently, we switched to a different initialization scheme which would rely only on the nuclei channel. In addition, this approach offers the advantage of not requiring any specific labeling to perform our brain alignment paradigm. This

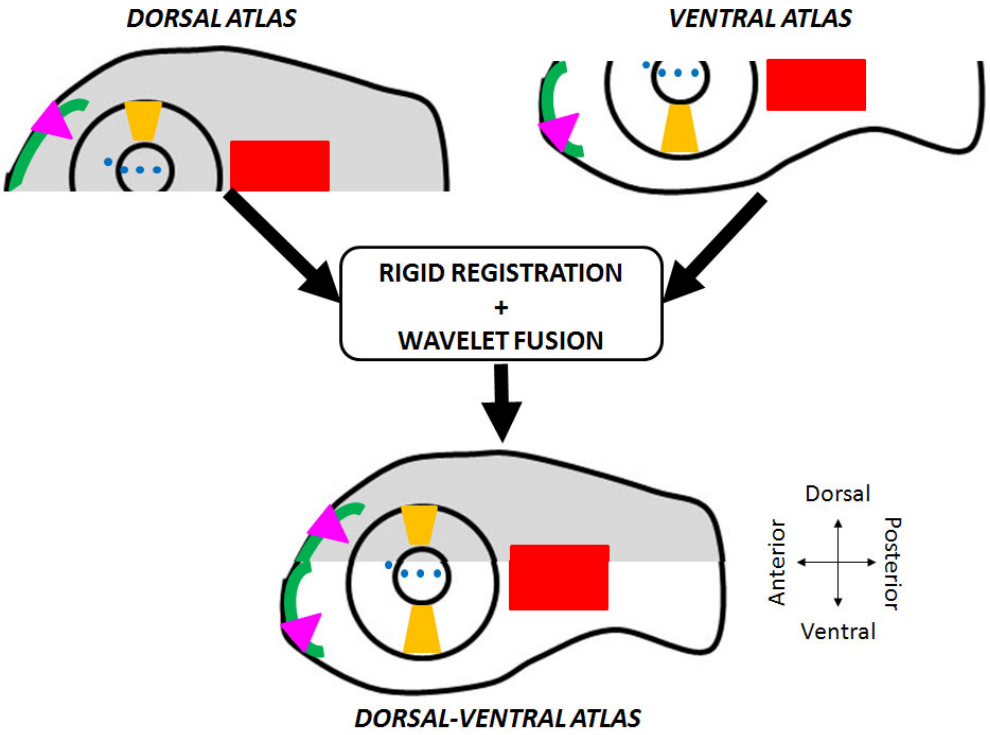

Figure 4.4: Schematic illustration of the dorsal-ventral integration.

way, the *th* marker does not have to be systematically acquired for all processed datasets, therefore freeing this channel for other purposes.

In particular, we manually selected 6 distinct landmarks per dataset (see Fig. 4.5): two of them were set to be the points where the anterior and posterior fibers split (F_A and F_P respectively), the other 4 corresponded to two couples of points placed at the foremost ventral and dorsal parts of the dataset and distributed along the anterior-posterior midline (M_1, M_2, M_3 and M_4 respectively). The manual selection of these landmarks was performed in ITK-SNAP (Yushkevich et al. [2006]) and took approximately 5 minutes per dataset.

Based on these 6 landmarks, we can extract a common referential defined by a three-vector basis comprising the anterior-posterior direction, $\overrightarrow{a}$, the dorsal-ventral direction, $\overrightarrow{d}$, and the lateral direction, $\overrightarrow{l}$. The two fiber-splitting points allow us to set $\overrightarrow{a}$. Fitting a plane (AP) to the two anterior-posterior couples lets us define $\overrightarrow{l}$, whereas $\overrightarrow{d}$ is the vector perpendicular to the previous two. The origin of this referential is placed at the posterior splitting fiber landmark. By

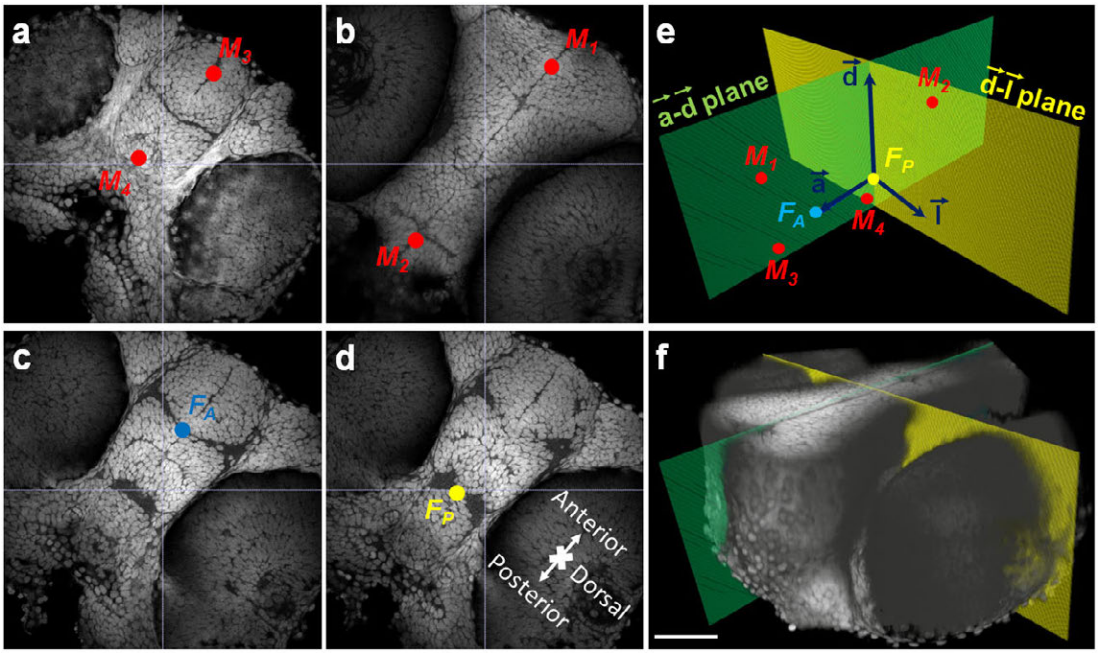

Figure 4.5: The 6 manually-selected landmarks used for the initial global alignment: (**a**) Ventral midline couple (M_3 and M_4, in red), (**b**) dorsal midline couple (M_1 and M_2, in red), (**c**) anterior fiber split (F_A, in blue) and (**d**) posterior fiber split (F_P, in yellow). (**e**) These landmarks are employed to extract the $(\vec{a}, \vec{d}, \vec{l})$ referential triplet (in dark blue). (**f**) View of the planes defined by vectors $\vec{a}$- $\vec{d}$ (in green) and $\vec{d}$ - $\vec{l}$ (in yellow) superimposed to the specimen nuclei channel (in gray). Scale bar 100 μm.

rigidly aligning the $(\vec{a}, \vec{d}, \vec{l})$ tuple from different zebrafish brain acquisitions we get a first, coarse alignment among them, see Fig. 4.6a.

4.3.2 Rigid vs. affine pre-alignment

We used the registration tools provided by the Advanced Normalization Tools (ANTs) platform in order to refine further the initial alignment between the brains of the *individual embryos* and the *template* prior to applying deformable transformation models on them.

In particular, we tested multi-scale rigid and affine registration models which optimized metrics based on mutual information and cross-correlation. Using an affine transformation with a cross-correlation metric, after substraction of the local mean from the image, yielded visually satisfactory results (Fig. 4.6b.) and

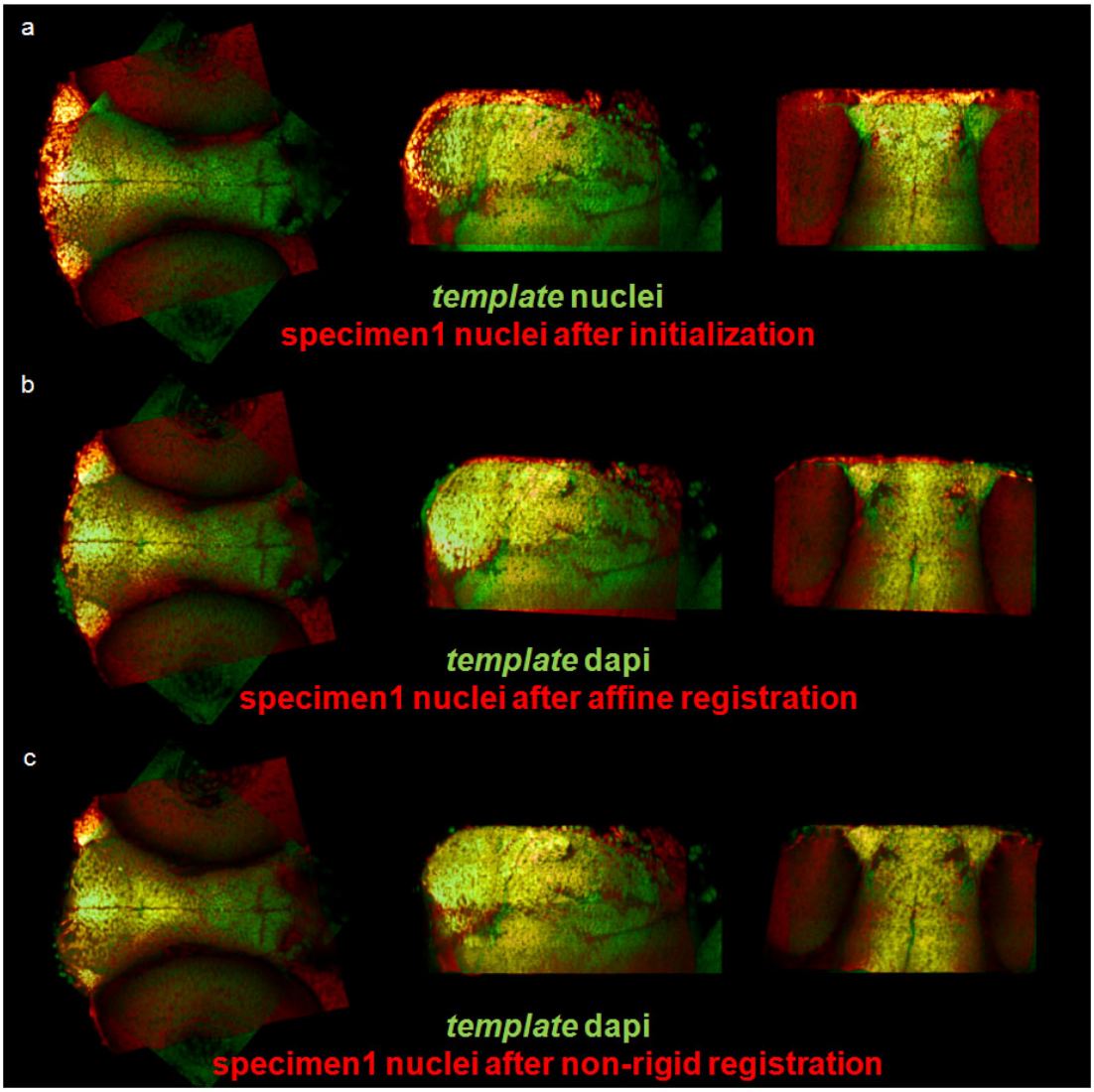

Figure 4.6: Alignment results between the brain *template* (in green) and one *individual embryo* (in red) after each of the workflow steps comprising an initial alignment based on the position of 6 predefined landmarks (**a**), an affine registration (**b**) and a non-rigid warp based on the nuclei channel (**c**).

scored well in our quantitative evaluation benchmark (section 4.4.2).

4.3.3 Non-rigid registration

After the global alignment and affine registration steps have brought the individual embryos into close alignment to the *template*, we apply a non-rigid registration

scheme. This deformable transformation model, guided by the anatomical information contained in the nuclei channel, deforms the analyzed specimen's structures to make them fit those of the *template*. Contrary to previous approaches (Peng et al. [2011], Ronneberger et al. [2012]), we chose not to employ landmarks and Thin Plate Splines to perform this non-rigid registration. We opted for a grayscale, non-rigid approach in an effort to make our brain alignment workflow more easily generalizable to new dataset and acquisition schemes. We considered that the nuclei channel itself conveyed sufficient information for the non-rigid registration to accurately bring into line all the relevant elements.

In particular, we perform a nonlinear transformation model called Symmetric Normalization (SyN) (Avants et al. [2008]), which has been shown to perform accurately and robustly in MRI human brain images (Klein et al. [2009], Avants et al. [2011]), see Fig. 4.6c. This transformation model has the advantage of being diffeomorphic and therefore invertible. The parameters employed include a gaussian regularization with different kernel sizes, a coarse-to-fine strategy in 3 different scales and the use of metrics based on mutual information and cross correlation (section 4.4.2). An example of the typical non-rigid warp applied to the datasets can be observed in Fig. 4.7.

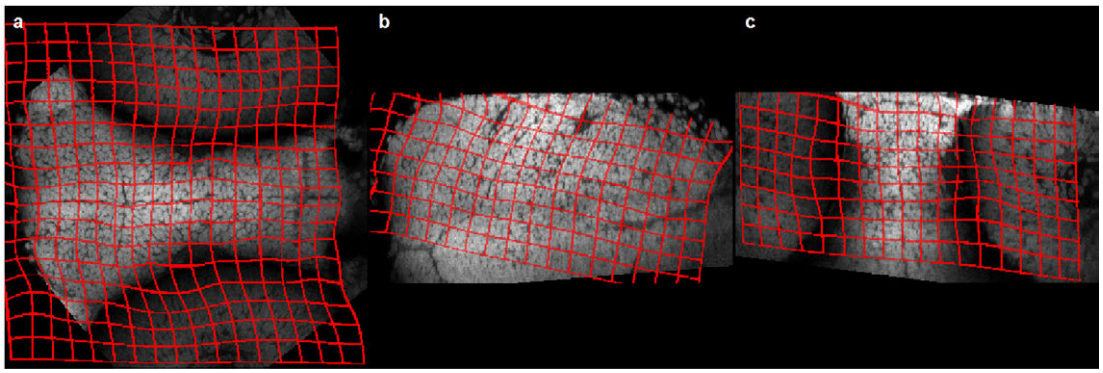

Figure 4.7: Example of the non-rigid warp underwent by an *individual embryo* to fit the *template*. (**a**) Transversal, (**b**) coronal and (**c**) sagittal orthoslices.

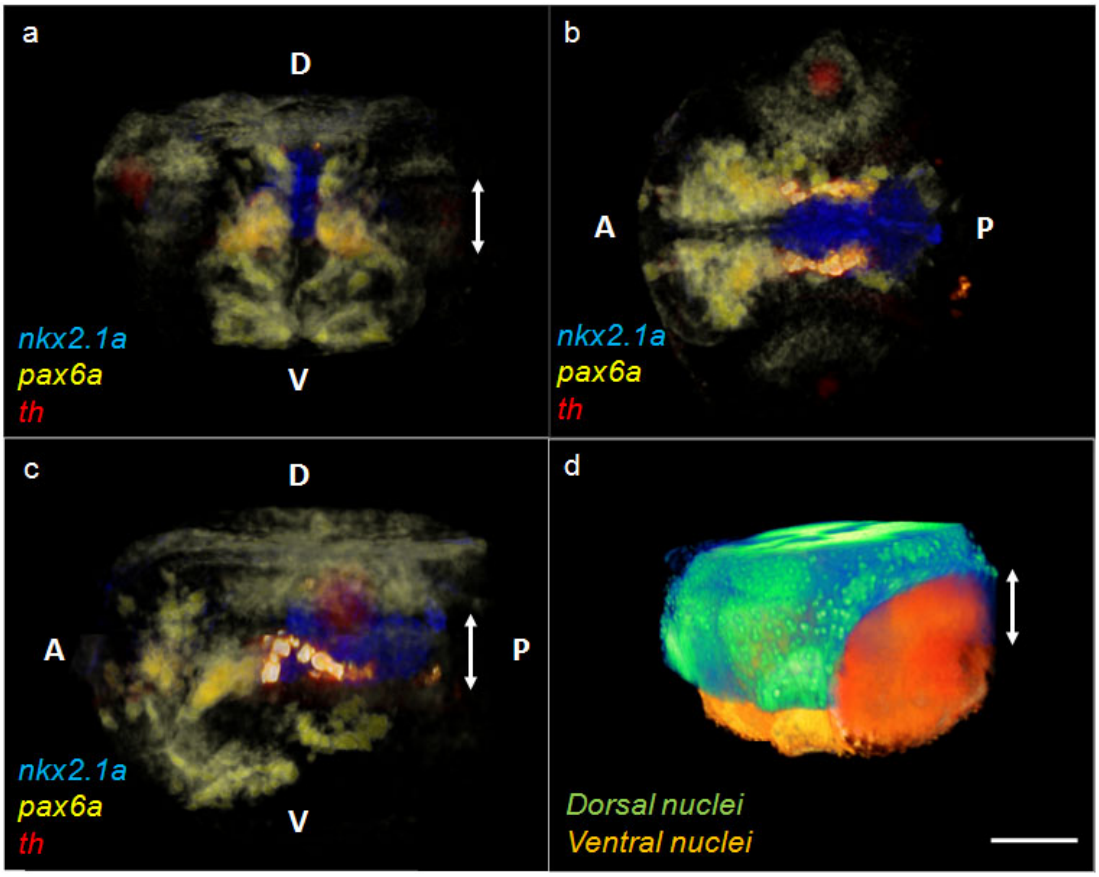

Figure 4.8: Fusion between the ventral and dorsal acquisitions of the 48 hpf brain *template*. (**a**) Frontal view, (**b**) dorsal view, (**c**) lateral view and (**d**) volume rendering of the nuclei and gene expression channels. The arrow indicates the region of dorsal-ventral overlap. Scale bar, 100 μm.

4.3.4 Fusion of a ventral and dorsal atlas

The steps described in sections 4.3.1-4.3.3 are separately applied to ventral and dorsal acquisitions in order to register them into their corresponding ventral or dorsal *template*.

Given that the dorsal and ventral *templates* are in fact acquisitions coming from the same individual, it is possible to find a transformation between them in order to combine both ventral and dorsal atlases into a whole brain atlas.

This transformation is obtained by applying a rigid registration algorithm between the ventral and dorsal views of the aforementioned *template*, followed by

a wavelet-based fusion scheme (Rubio-Guivernau et al. [2012]) that selects the best-detailed features from both images in order to compensate for the depth penetration losses, see Fig. 4.8.

4.4 Validation

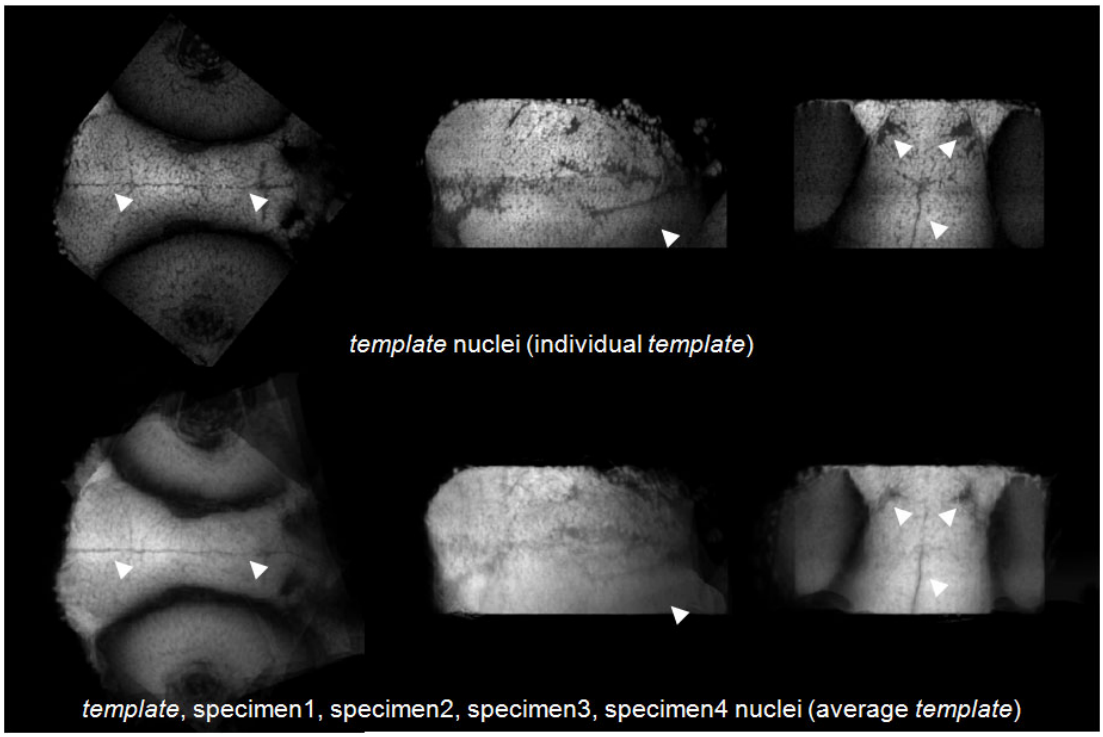

Figure 4.9: Nuclei of the individual *template* employed (**a**) and the result of averaging the nuclei of the five processed datasets: the individual *template* plus the four aligned *individual embryos* (**b**). The most prominent morphological traits (see arrowheads) can be observed in both the individual and the average template indicating that the non-rigid warp successfully managed to align them.

4.4.1 Qualitative evaluation

As a way to qualitatively evaluate the registration performance, we averaged the nuclei channels of the 5 aligned ventral brains at 48 hours. In Fig. 4.9 we can

appreciate how the average *template* keeps showing the same anatomical traits as the individual *template* implying that those parts were properly aligned.

Another qualitative evaluation test was performed for a particular dataset: Fig. 4.10 shows how gene expression *nkx2.1a* is warped by the non-rigid registration scheme in that particular case so that it follows the same anatomical location in the *template* as it did in the original *individual embryo*.

4.4.2 Quantitative evaluation

A quantitative evaluation of our non-rigid registration scheme is a fundamental step in order to optimize its performance (Lombardot et al. [2008]). We used the ventral atlas at 48 hours as the benchmark where to run quantitative evaluation tests. These tests consisted in measuring how well certain biologically-relevant structures, annotated in both the individual brains and the *template*, are aligned after the non-rigid warp according to certain evaluation criteria.

Manual segmentations of 3 relevant structures were annotated using ITK-SNAP and supervised by expert biologists (Fig. 4.11):

- The forebrain outer border (BOR).
- The cross-shaped ventricles boundaries (CRO).
- The forebrain fibers (FIB).

These 3 structures were labeled both on the *template* and on three individual brains. Different evaluation measures were considered to measure the distances between each of the 3 annotated structures in the *template* and those in the registered individual datasets:

- Maximum Haussdorf distance (H_M).
- Mean Haussdorf distance (H_m).
- Volume overlap measured by the Jaccard index ($J_{A,B} = \frac{|A \cap B|}{|A \cup B|}$).

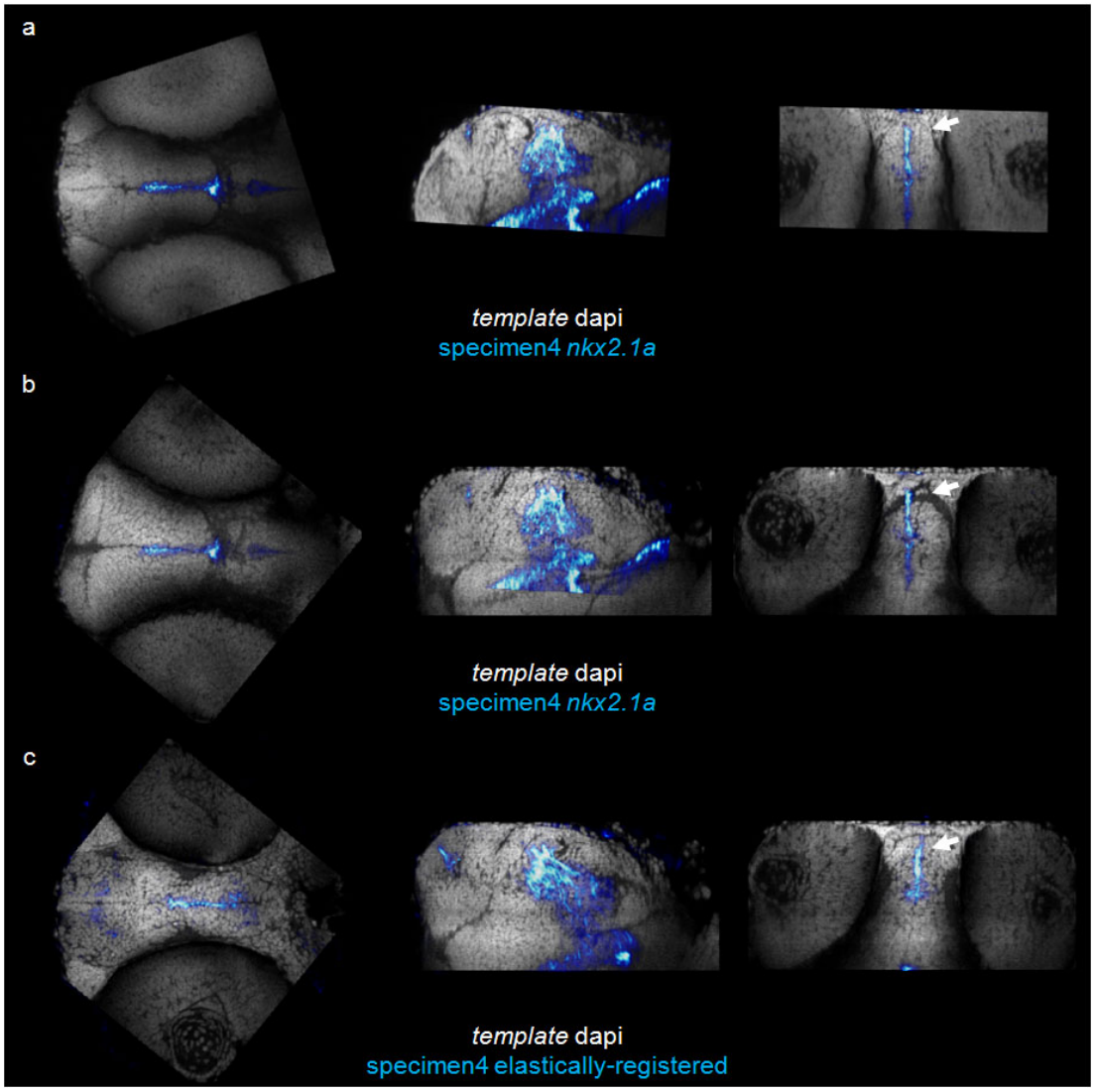

Figure 4.10: (**a**) Shows gene expression *nkx2.1a* together with the nuclei of the *individual embryo* where it was acquired. We can note this gene pattern is distributed along the morphological border between the brain lobes. (**b**) Shows the same gene expression superimposed to the *template* nuclei after having undergone the initialization and affine registration steps. We can note that the affine-registered gene expression does not match the equivalent morphological border in the *template* whereas the non-rigid warp solves this problem (**c**).

The maximum Haussdorf distance was only used for the forebrain outer borders (BOR) in the individuals and *template*. Regarding the cross-shaped ventricle boundaries (CRO) and the forebrain fibers (FIB), just the Jaccard index (J) and

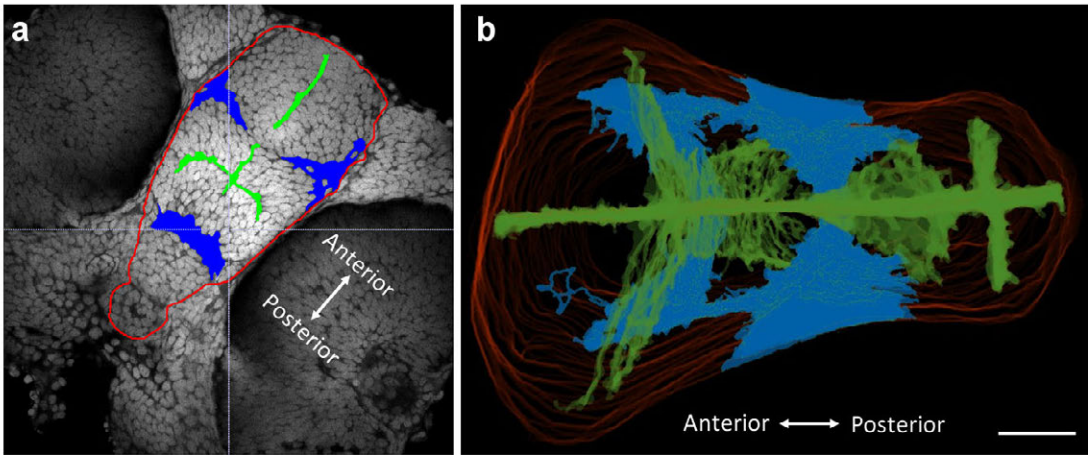

Figure 4.11: (**a**) Transversal orthoslice displaying the manual segmentations of the 4 biologically relevant structures considered for the method evaluation: brain outer border (in red), ventricules border (in green) and fibers (in blue). (**b**) Volume rendering of those annotated structures.

mean Haussdorf (H_m) distances were considered. In these cases, the complexity and biological variability of these structures make the computation of the distance between exactly equivalent points an ill-defined problem, while calculating global measurements appears to be a more reliable evaluation. Because of that, when measuring the maximum Haussdorf (H_M), unstable and spurious measurements that are not representative of the global behavior of the alignment algorithm might appear. Therefore the mean Haussdorf distance (H_m) was preferred as it provides a global measurement of distance

In order to optimize the results of our mapping framework and select the best parameters, different registrations were run based on different parameters sets. After each execution, the evaluation measures between the deformed annotated structures from the individual embryos and those of the *template* were calculated. The parameters to configure during this evaluation tests were:

- The registration metric. Two different metrics were tested: Mutual information (MI) and cross correlation after substraction of the local mean from the image (PR).

- The transformation model. A nonlinear transformation model called Symmetric Normalization (SyN) (Avants et al. [2008]) was employed with dif-

ferent optimizer steps sizes. A smaller step size is generally more accurate but takes more time to compute and may not capture as much deformation if the optimization gets trapped in a local minimum.

- The regularization term. A Gaussian regularization, with different kernel sizes, was used to smooth the resulting deformation field in every iteration.
- Multi-resolution scheme. Three multi-resolution levels were employed concentrating most of the iterations on the two most downsampled and, therefore, less computationally-heavy scales.

The best performance was achieved using the PR cross correlation metric with substraction of the local mean from the images using a 5 μm radius, the SyN transformation model with an optimizer step of 0.5 μm, 3 multi-resolution levels and a 100 μm Gaussian regularization. Only in specimen 3 results were better using a regularization weight of 150 μm.

Tables 4.1-4.3 show the results after applying our non-rigid registration scheme confirming the improvement in all the evaluation metrics compared to the initial global alignment or to an affine registration scheme (Fig. 4.12). Indeed, the affine pre-alignment did not always improve the results compared to the initialization. Future refinements include dealing with border effects as they were generally responsible for the maximum Haussdorf distances (H_M) detected between the forebrain borders (BOR).

This set of parameters, configured from our 3 benchmark cases, were used to run the rest of registration instances (Fig.4.14).

4.5 Results: A gene expression atlas of the 48 hpf zebrafish brain

Applying the methodology described in section 4.3 and tuned in section 4.4.2, we reconstructed a prototype of the dorsal and ventral atlas at 48 hours.

In the ventral side, we employed five datasets: One *template* and four different individual embryos which were registered onto it, see Fig. 4.13. The ventral atlas is then composed by eight different gene expression patterns: *nkx2.1a*, *pax6a*,

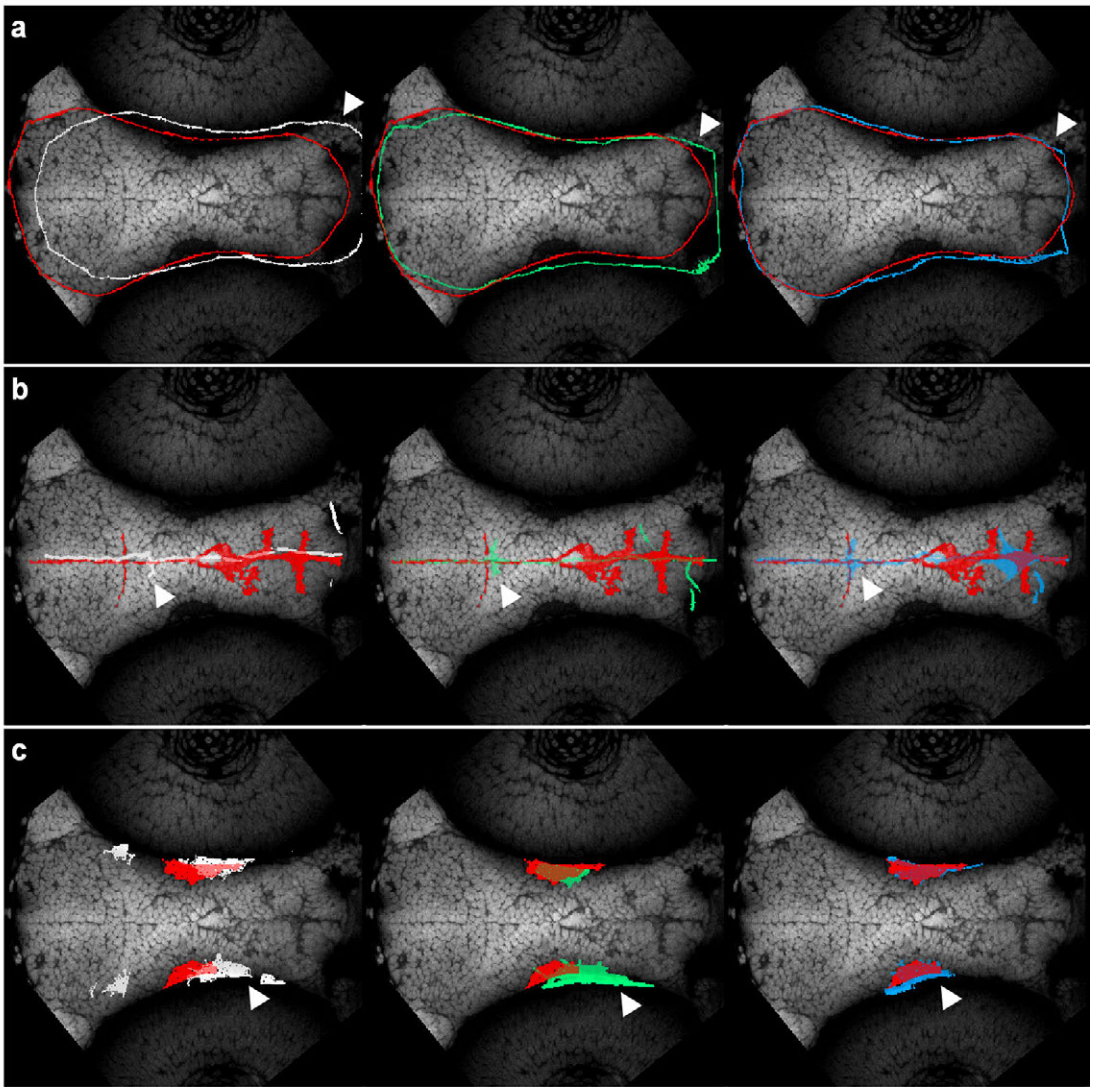

Figure 4.12: The forebrain border (*BOR*) (**a**), the cross-shaped ventricles (*CRO*) (**b**) and the forebrain fibers (*FIB*) (**c**) of the *template* (in red) compared to their corresponding structures in one specimen after initialization (left column, in white), affine registration (center column, in green) and non-rigid registration (right column, in blue). White arrowheads point out areas with clear improvements of our non-rigid approach compared to the initialization and affine schemes.

dlx2a, *sim1*, *tbr1*, *rx3*, *six3b* and *dbx1a*. The fact that two of these gene expressions, *nkx2.1a* and *dlx2a*, were repeated in different acquisitions will allow to perform a study on gene expression inter-subject variability. A qualitative display of this atlas can be observed in Fig. 4.14.

dataset	After initialization	After affine	After non-rigid	metric
specimen1	10.03 μm	5.10 μm	3.15 μm	
specimen2	13.94 μm	9.75 μm	7.13 μm	H_m
specimen3	11.25 μm	6.62 μm	3.99 μm	
specimen1	34.35 μm	32.14 μm	23.38 μm	
specimen2	35.66 μm	33.41 μm	30.80 μm	H_M
specimen3	31.17 μm	25.58 μm	20.87 μm	
specimen1	0.73	0.87	0.92	
specimen2	0.56	0.74	0.81	J
specimen3	0.65	0.79	0.88	

Table 4.1: Evaluation metrics for the forebrain borders (BOR) after initialization, affine pre-alignment and non-rigid registration.

dataset	After initialization	After affine	After non-rigid	metric
specimen1	4.65 μm	2.76 μm	1.50 μm	
specimen2	5.54 μm	5.55 μm	2.91 μm	H_m
specimen3	1.90 μm	3.87 μm	1.73 μm	
specimen1	0.06	0.17	0.30	
specimen2	0.02	0.07	0.10	J
specimen3	0.08	0.07	0.22	

Table 4.2: Evaluation metrics for the cross-shaped ventricles (CRO) after initialization, affine pre-alignment and non-rigid registration.

dataset	After initialization	After affine	After non-rigid	metric
specimen1	7.83 μm	5.49 μm	3.82 μm	
specimen2	5.95 μm	9.23 μm	6.12 μm	H_m
specimen3	4.74 μm	5.45 μm	3.00 μm	
specimen1	0.15	0.31	0.54	
specimen2	0.14	0.09	0.24	J
specimen3	0.13	0.19	0.45	

Table 4.3: Evaluation metrics for the forebrain fibers (FIB) after initialization, affine pre-alignment and non-rigid registration.

In the dorsal side, we employed two datasets: One *template* and one individual embryo. The dorsal atlas was then composed by four different gene expressions:

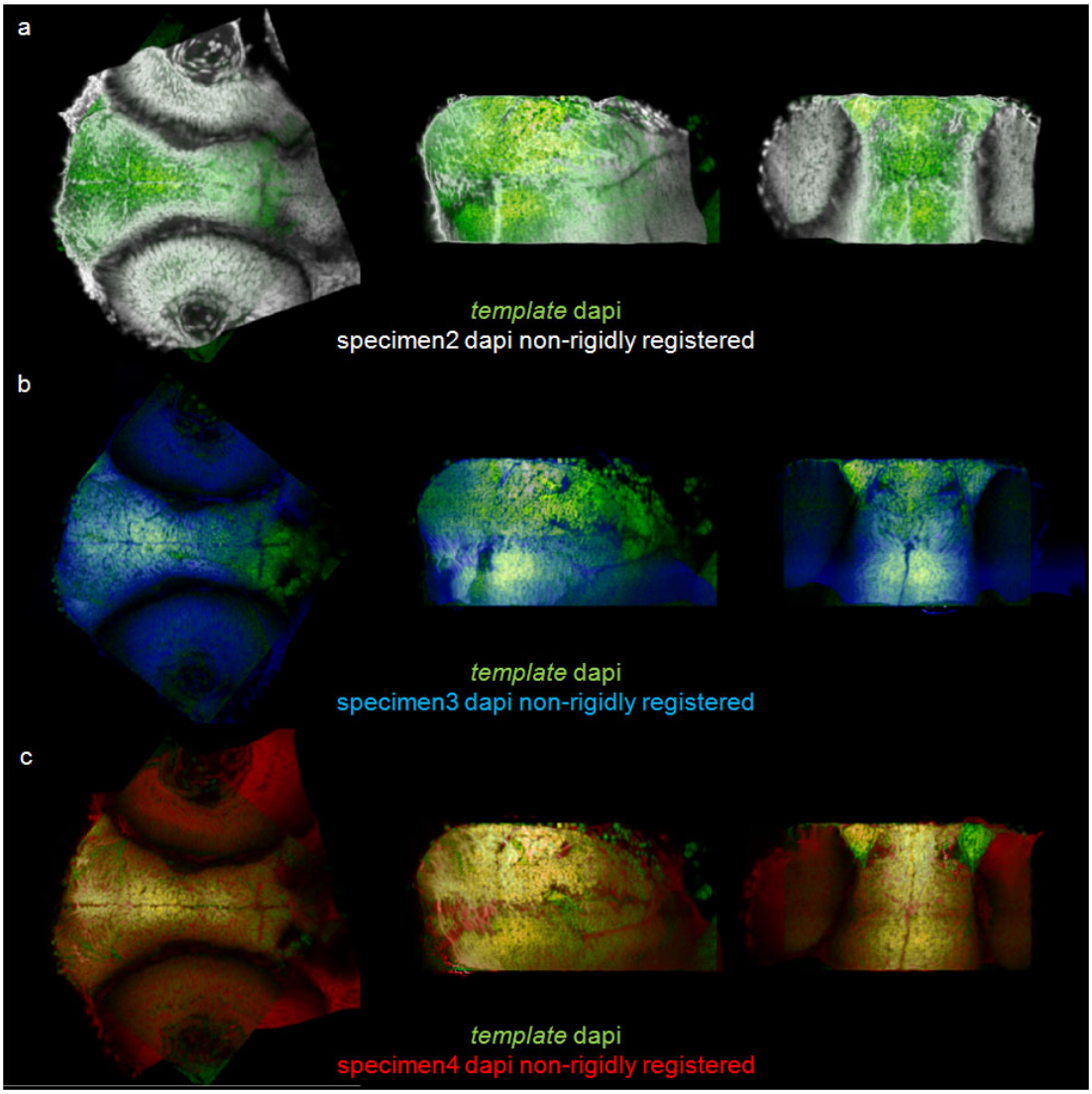

Figure 4.13: The presented workflow achieves a final alignment between the brain *template* (in green) and the *individual embryos* (**a-c**).

nkx2.1a, *pax6a*, *dlx2a* and *tbr1*.

This atlas can be exploited by querying which gene patterns express in which regions. For this purpose, a segmentation of the atlas *template*, as proposed by Ronneberger et al. [2012] (Fig. 4.15a), could provide its regionalization into distinct anatomical parts.

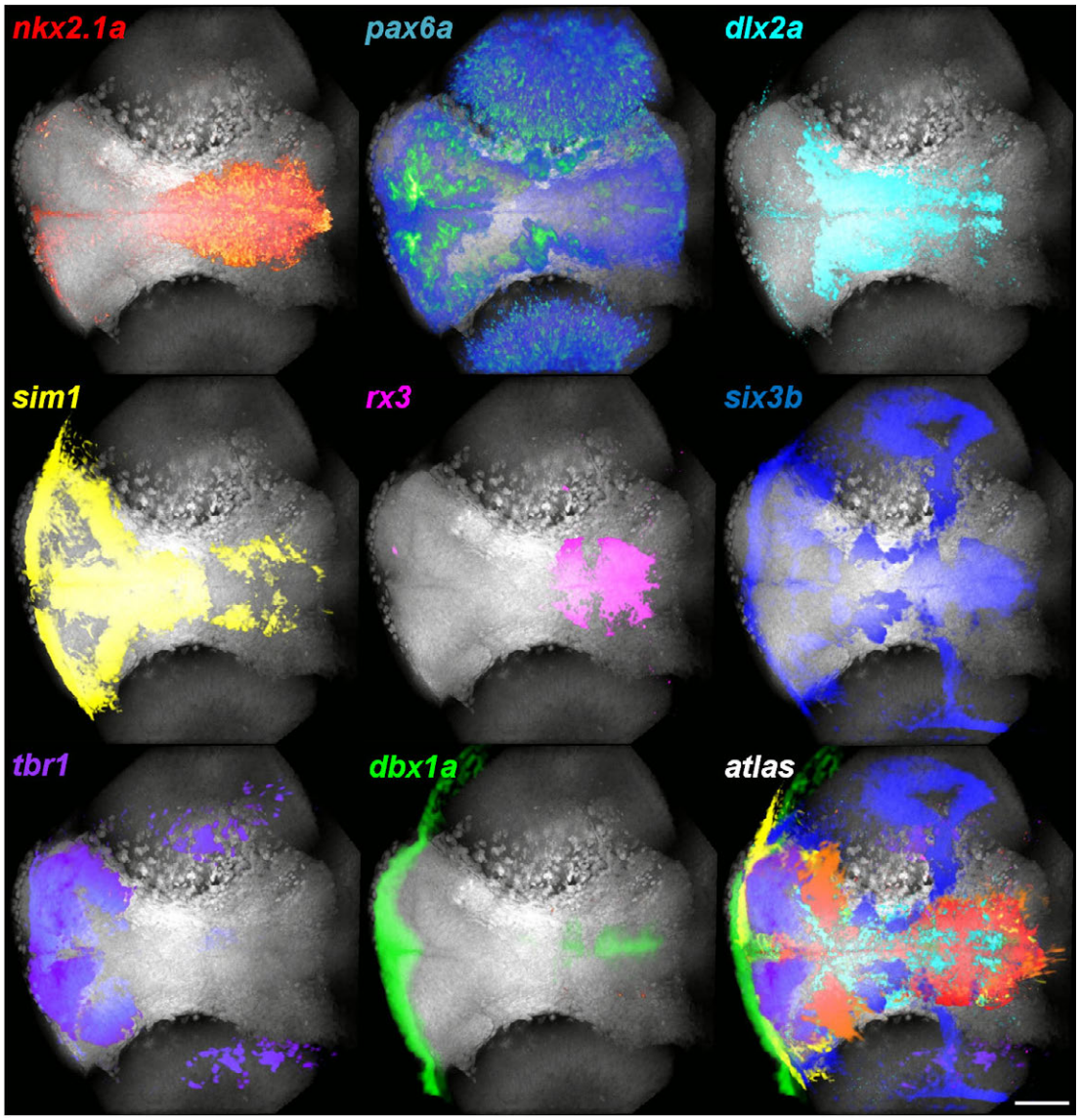

Figure 4.14: Dorsal view of the eight aligned gene expressions, coming from the five employed *individual embryos*, superimposed to the *template* nuclei (in white) at 48 hpf. Scale bar 100 μm.

4.6 Discussion

We have designed and implemented a novel framework that specifically tackles the gathering of gene expression data in late embryo brain development, a period where the appearance of distinct morphological landmarks requires the use of

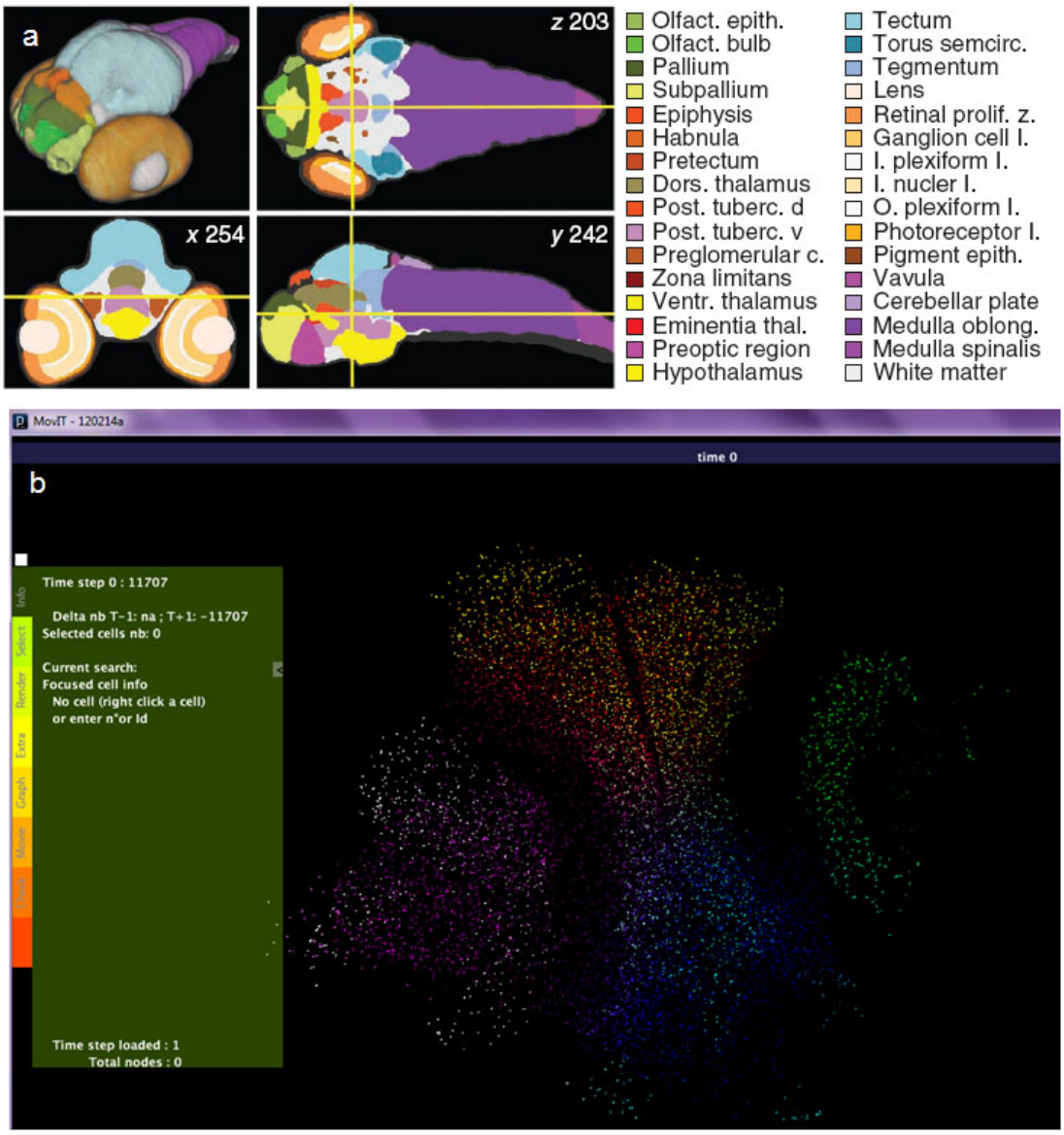

Figure 4.15: (**a**) State-of-the-art atlases of gene expression in the brain limit co-expression queries to tissular-level regions defined manually in the *template* according to anatomical divisions (Ronneberger et al. [2012]). (**b**) Nuclei detection performed on the ventral *template* dataset at 48 hours.

non-rigid transformation models to match different individuals into a common template. Depth signal extinction, another common limitation to late development studies, has been avoided by separating the atlasing strategy into two different views, ventral and dorsal, which are processed separately to be later

registered and fused together. In addition, this framework has been designed to operate even with the most basic data acquisition procedure: no specific referential labeling is required for the registration procedure which relies only on the nuclei staining channel and some initialization landmarks.

Even though a common strategy in literature (Peng et al. [2011], Ronneberger et al. [2012]) consists in creating an average *template* (see Fig. 4.9), we opted to employ an individual specimen as the *template* where to register the rest of acquisitions. Although average *templates* have certain advantages, they display smooth shapes and contours where other individuals can more easily fit, it would not be possible for them to link the resulting atlas with a particular cell lineage. This is the main reason why we decided to employ an individual *template* as we eventually expect to make the cell lineage link with our atlas at earlier stages to characterize the formation and regionalization of the hypothalamus and first *th* positive neurons.

In future developments, in order to have the advantages from both approaches, the individual *template* should be selected to be the specimen of the cohort that minimizes the transformations required to map the rest of the population. Such a prototypic individual would additionally allow working with a *template* for which it is possible to extract a cell lineage.

So far, no real cellular level atlas of an organ at late developmental stages have been released. Generally, this kind of atlases are limited to the tissular resolution, that is to say, gene co-expressions can be only studied in some previously annotated anatomical domains (Fig. 4.15a). We tested a nuclei detection algorithm (Zanella et al. [2010]) on our images together with the interactive visualization tool MovIt (Savy and Bioemergences [2013]) which allows to visually modify and select the best set of parameters to obtain a satisfactory nuclei detection in the developing brain. The visual inspection of the results revealed that nuclei detection was far from perfect in such a demanding scenario, where nuclei are highly packed together (Fig. 4.15b). However, further work will be carried out in this sense, including testing new cell extraction algorithms specialized in dense tissue environments (Pop et al. [2013]). This strategy could open up the possibility to perform multicellular-level gene expression queries to automatically characterize the groups and dynamics of gene co-expression (section 3.7).

Future work includes linking the resulting zebrafish brain gene expression resource to a powerful interactive visualization interface such as Atlas-IT to promote the use of our strategies. Another future development may include setting up automatic landmark recognition schemes (Stern et al. [2011], Ronneberger et al. [2012]) that can systematically recognize the morphologically distinct points required for initialization.

I will pause to consider this eternity from which the subsequent ones derive.

Jorge Luis Borges

Chapter 5

A framework to link gene expression data and cell lineage from early to late zebrafish development

5.1 Introduction

Understanding the role of gene expression in cell dynamics and morphogenetic processes during the embryogenesis of living animals is a major challenge in biomedical research (Megason and Fraser [2007]). Capturing quantitative data of gene expression in time and space at the cell resolution level becomes a crucial step to fulfill this goal.

Achieving 3D quantitative genetic data at the cellular level has been approached by a few big projects (Lein et al. [2007], Fowlkes et al. [2008], Long et al. [2009]) concerning different animal models: the mouse brain, the *Drosophila* or the *C. elegans*.

Although the state of the art in gene expressions atlases is still 3D, we can envision the extension of such strategies to 3D+time as a promising way to combine temporal and spatial information, nuclei trajectories with genes levels of expression, so that correlations between gene dynamics and cell dynamics can be

explored.

Concerning the zebrafish (*Danio rerio*) model, the construction of a spatiotemporal, 4D (3D+time) map of gene expression in zebrafish development has challenging characteristics: at early developmental stages there are no clear anatomical references, the whole development is not stereotyped at cellular level as it is in *C. elegans*, and its embryogenesis complexity (cell morphology, morphogenetic movements, etc.) is higher than in the *Drosophila* case.

Previous works on zebrafish include strategies to either create a 3D map of gene expression levels (Castro et al. [2009], Ronneberger et al. [2012]) or to reconstruct an *in toto* global description of cell dynamics (Keller et al. [2008], Olivier et al. [2010]). However, these two separate concepts had never been previously combined.

In this chapter, we first introduce a proof-of-concept experiment which integrates a quantification of the gene expression products and their cellular location together with a 3D+time (Bao et al. [2006]) cell tracking (section 5.4). As a result, we get a complete (x, y, z, t, g) description of each embryo cell (Megason and Fraser [2003]). Additionally, we also introduce a preliminary attempt to construct a 3D+time atlas by registering 3D gene expressions into an *in vivo* acquisition of a living specimen (section 5.5). As a result, we get a complete $(x, y, z, t, g_1, ..., g_N)$ description of each embryo cell. These strategies merge gene expression data with the lineage tree of a living specimen and allow to correlate gene pattern combinations with cell fate. With simple modifications, this systematic analysis could be easily extended to other developing stages and genetic expressions.

To this end, we employed time-lapse, confocal, bi-photon laser scanning microscopy with the aim of observing the spatial location and *in vivo* evolution of zebrafish nuclei during the gastrulation period (Kimmel et al. [1995]) from 6 to 9 hours post fertilization (hpf). During this period, the transcriptional factor we focus on -named *goosecoid* (*gsc*)- starts its expression at the dorsal side to progressively migrate up to the animal pole leading to the dorsal axis formation (Fig. 5.1).

This chapter is organized as follows: section 5.2 depicts the different datasets employed. Section 5.3 shows the general framework and describes the computational methods devoted to cell detection, cell tracking, gene quantification and *in*

situ to *in vivo* data registration. These prospective methodological tools constitute a step towards studying how gene expressions influence cell dynamics and differentiation. To this end, we set up two different experiments: 1) In section 5.4, we started by dealing with one transgenic line, that is, an *in vivo* dataset stained to reveal the live evolution of both one gene product together and the embryo nuclei. The application of gene expression quantification together with cellular tracking techniques constituted the first tentative in literature fusing lineage and gene expression dynamics, both in terms of global cell behavior and particular migration of 10 representative cells. 2) Section 5.5 describes an introductory framework towards the integration of further gene expression products so that a more comprehensive study about the influence of synexpression groups on cell dynamics can be overtaken. In this sense, we implemented a tentative workflow that matches *in situ* gene expressions resources (3D atlases) into an *in vivo* 3D+time specimen. Finally, we discuss the results and future work related to this line of research.

5.2 Dataset description

We employ two different types of microscopy data in this chapter: One *in vivo* dataset depicting a complete living embryo and several *in situ* datasets depicting embryo portions fixed for a particular developmental stage.

The *in vivo* image dataset was acquired by a bi-photon confocal laser scanning microscope Leica SP5 with a 10X objective and 1030nm and 980nm laser light excitation which resulted in two differentiated channels: one holding the *egfp* expression in the zebrafish (a live reporter of gene *goseecoid* (*gsc*)), the other holding the embryo nuclei. We developed a new GFP transgenic line to label the *egfp* expression, while the nuclei were labeled with a H2bmCherry mRNA injection. The resulting 3D stacks are composed of a set of 2D images, homogeneously spaced in depth, acquired in the XY plane from the animal pole. Image size was 512x512x245 voxels with a voxel resolution of 1.51x1.51x1.51 μm. The repetition of this imaging process once every 2m44secs generated a sequence in time depicting the embryo *in vivo* development between 6 and 9 hpf (Fig. 5.1).

The *in situ* image datasets are the same as those employed for the construction

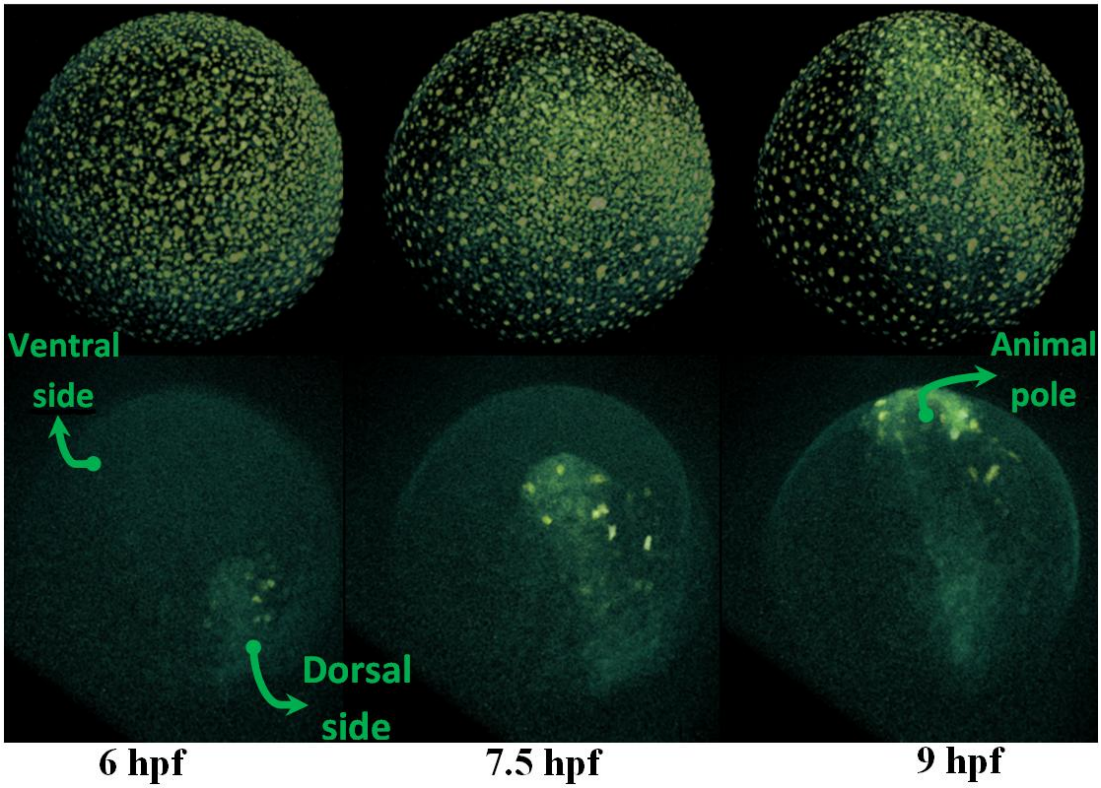

Figure 5.1: Depiction of a live zebrafish embryo acquired *in vivo*. Nuclei (top) and *egfp* (bottom) channels are shown at 6, 7.5 and 9 hpf.

of the 3D gene expression atlas resource (see section 3.2). In particular, we make use of the embryo cohorts fixed for three different developmental stages at 5.3, 6 and 6.3 hpf. These embryos were also acquired by a bi-photon confocal laser scanning microscope and labeled to reveal three channels: one holds the embryo nuclei whereas the two others hold two different *in situ* gene expressions, one of which it is always the referential *gsc* pattern.

5.3 Methodology

We propose a computational framework to quantify cellular position and gene expression levels throughout early zebrafish embryogenesis captured over a time-lapse series of *in vivo* 3D images.

Our strategy involves nuclei detection, cell geometries extraction, automatic

gene levels quantification and cell tracking to reconstruct cell trajectories and the lineage tree. We also include a strategy to register *in situ* gene expressions into *in vivo* embryos. Each cell in the embryo is then precisely described at each given time t by a vector composed of the cell 3D spatial coordinates (x, y, z) along with its gene expression level g. This comprehensive description of the embryo development is used to assess the general connection between genetic expression and cell movement. We also investigate genetic expression propagation between a cell and its progeny in the lineage tree.

5.3.1 Nuclei detection and tracking

We performed a cell nuclei detection based on the numerical solution of a 3D nonlinear advection-diffusion equation proposed by Drblikova et al. [2007]. Given the set of identified nuclei, Melani et al. [2007] implemented an iterative greedy algorithm that builds the cell lineage tree as follows: The vector field of every image I_t is computed by registering it into next image I_{t+1}. Based on this vector field, it is possible to predict the position of each detected nuclei at I_t in I_{t+1}, then assign the closest detected nuclei in $t + 1$ to be the continuation of its trajectory. Mitosis are identified by the Hough transform, which detects dividing nuclei -which are not longer spherical-, and by the presence at I_{t+1} of two nearest-neighbors. After visual inspection, the experts assessed the classification to have at least a 95% of correctly tracked cells per time step.

5.3.2 Cell lineage validation

We used the visualization tool Mov-It (Savy and Bioemergences [2013]) which was specifically designed to visualize and validate the described data. Its validation unit includes the possibility of correcting false positives (removing nuclei), false negatives (adding nuclei) and adjusting nuclei positions, Fig. 5.2. It is also possible to verify cells lineage by validating the links between their temporal trajectories, removing false mitosis, creating new links to manually added nuclei, etc.

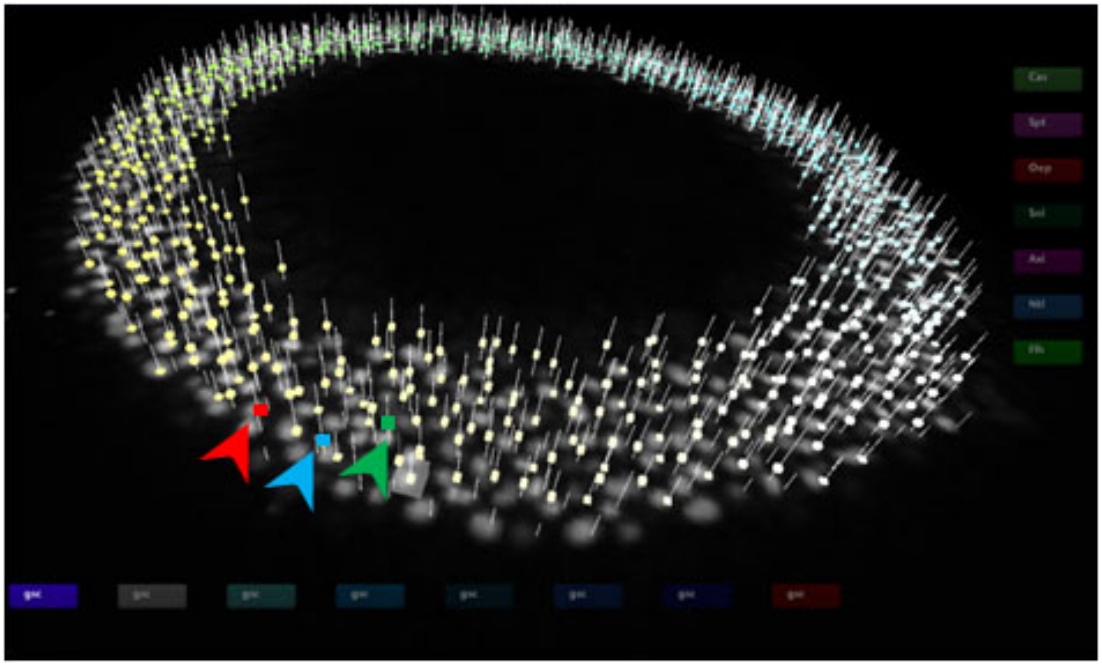

Figure 5.2: Validating nuclei detection and tracking with Mov-IT. Wrong nuclei detection and progeny links can be annotated (red arrowhead), corrected (green arrowhead) or added (blue arrowhead).

5.3.3 Cell geometries extraction

Given that our dataset did not dispose of membrane images, we used the detected nuclei centers as seeds to generate their corresponding Voronoi regions within the embryo volume. The Voronoi region associated to each nuclei n_i can be described as the set of points which are closer to n_i than to any other nuclei, (Fig. 5.3). Luengo-Oroz et al. [2008] showed that these Voronoi regions can provide an approximate model of the true cell geometries.

5.3.4 Automatic gene quantification

We assumed the gene expression level to be proportional to the fluorescence brightness within the embryo. Therefore, our approach was to use the mean intensity value of the *gsc* channel within each segmented cell in order to assign them a genetic activity score ranging from 0 to 1. However, there are many factors that can significantly bias these scores. In consequence, we performed two

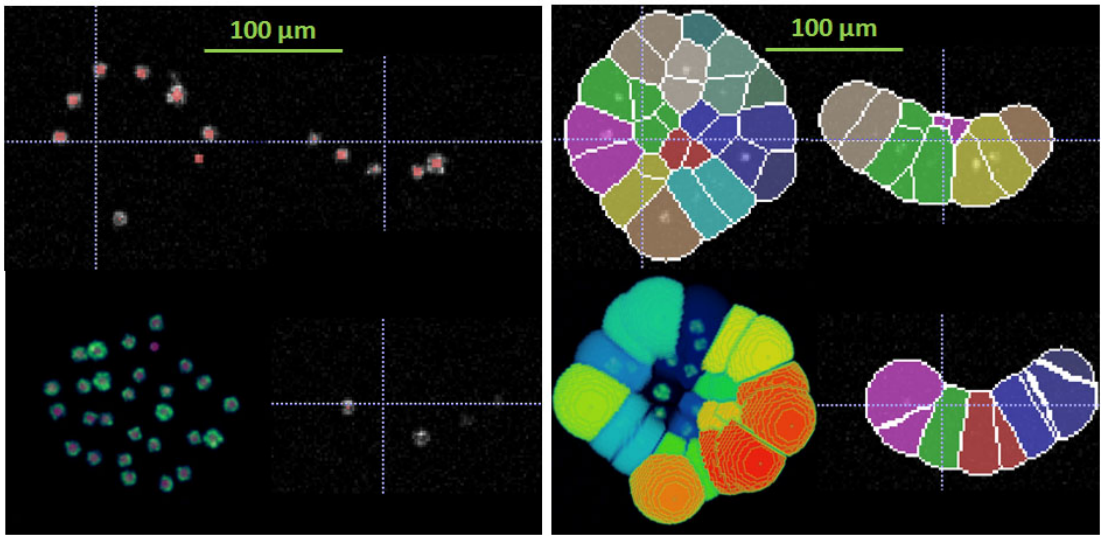

Figure 5.3: The left panel shows (from top to bottom, from left to right) the transversal, coronal, sagittal views and volume rendering of the raw nuclei (white) and detected centers (red) whereas the right channel displays the associated Voronoi segmentation.

kinds of normalization: first, in order to compensate for differences on the depth extinction and object transparency, we weighted our cell scores by the luminosity of their corresponding nuclei (which is supposed to be stable for every nucleus). Then, to enable comparison across the time-lapse acquisition, we applied an intra-step temporal normalization, meaning that values assigned to each cell were not absolute but dependent on how they rated against the rest of the cells present at their same time stage.

5.3.5 Registration of a 3D atlas into a 3D+time *template*

In chapter 3, we described a general workflow, Match-IT, that can achieve the alignment of any *analyzed specimen* specimen into a referential individual or *template*. In that case, both *analyzed specimens* and *template* were acquired *in situ*, that is to say, they were fixed to a particular developmental stage.

Here, we extend the Match-IT workflow so that the *in situ analyzed specimens* employed in chapter 3 can be integrated onto a living *template* acquired *in vivo*, see Fig. 5.8a-b. First, for each of the 6 developmental stages to which the different *in situ analyzed embryos* belong, we identified a corresponding time-

lapse point in the live *template*. Then, we employed Match-IT to register each cohort of *analyzed specimens* to their corresponding time-lapse point in the live *template*. The nuclei and the *gsc* channel were used to guide the alignment procedure performed by Match-IT (section 3.3). However, in this case, we dispose of *in situ* gsc expression in the case of the *analyzed embryos*, but not in the *in vivo template* where a transgenic *egfp* line was revealed. Having previously assessed the correspondence at early developmental stages between *in situ gsc* expression and its corresponding *in vivo egfp* reporter (section 3.7.1), we could use this live transgenic line as the proper referential to guide, together with the embryo shape, the Match-IT registration algorithm.

5.4 Integration of gene expression quantification and cell tracking *in vivo*

This experiment, which aims at describing each cell by its position at the (x, y, z, t, g) space, includes two main schemes (Fig. 5.4): 1) An automatic genetic quantification scheme that provides us information about the levels of expression of one gene, *egfp*, and its location in time and space and 2) An automatic lineage reconstruction scheme that provides each embryo cell progeny and trajectory which can be visually validated. This methodology will be tested in the expression pattern of transcriptional factor *(egfp)* through the gastrulation process between 6 and 9 hours post fertilization (hpf). The Mov-IT visualization platform (Savy and Bioemergences [2013]) was employed to label and verify the trajectories of 10 clones (10 cells together with all their progeny) which were used to study in detail the evolution of cell speed and *gsc* expression throughout the cell lineage (sections 5.4.1-5.4.2).

5.4.1 Results: Statistical global patterns

Following the scheme depicted in Fig. 5.4 our gene quantification strategy was employed to obtain measures of the *egfp* levels of expression at every cell detected between 6 and 9 hpf. Likewise, our cell tracking strategy let us compute the current speed and direction of each cell during that period (Fig. 5.5). After

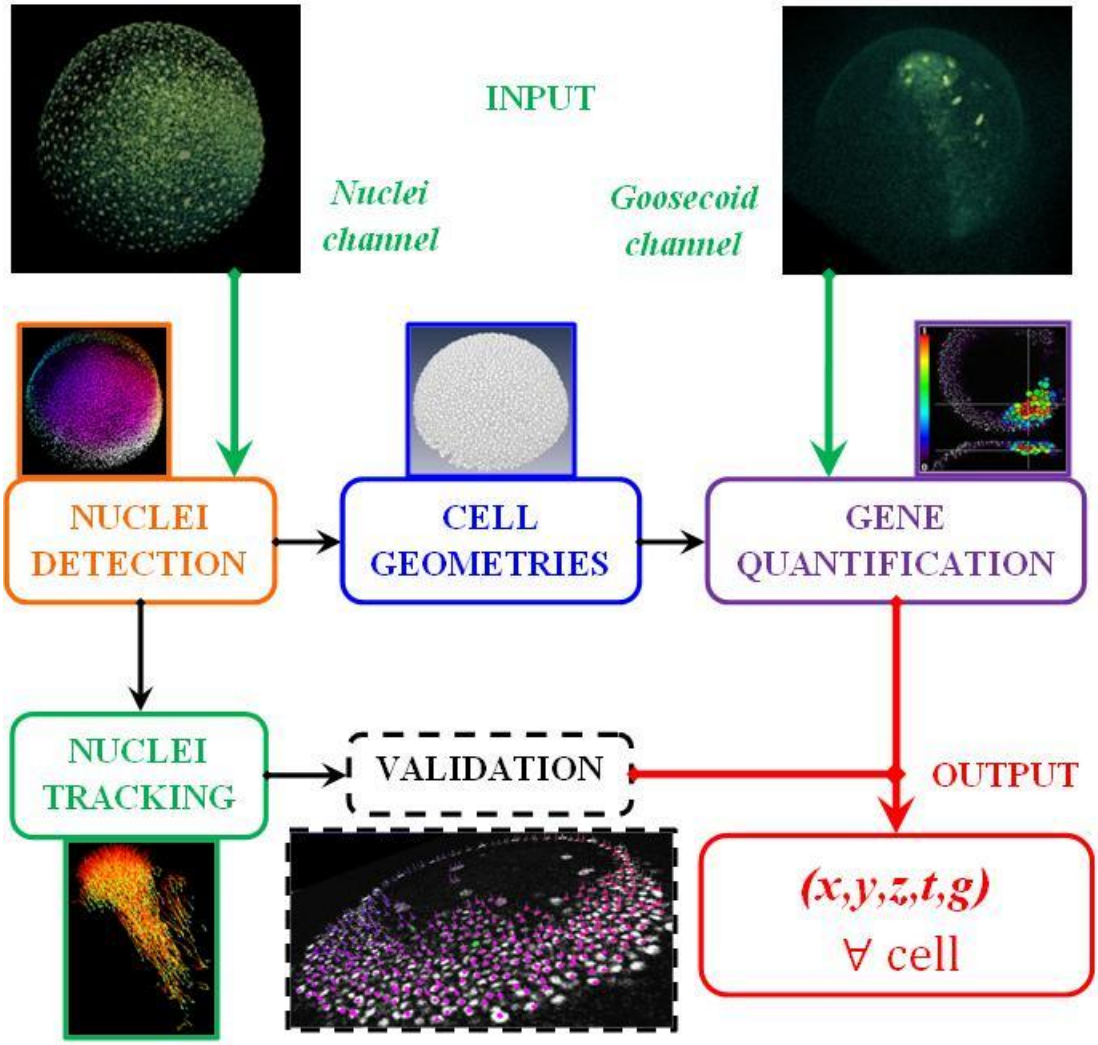

Figure 5.4: Block diagram depicting the *in vivo* gene quantification and cell tracking framework.

comparing both *egfp* expression and cell speed, we can observe (Fig. 5.5) that movements towards the dorsal side -the location where *egfp* starts its expression- begin with cells placed at the opposite ventral side which did not express *egfp* at first. However, as time goes by, we can appreciate there is an increasing connection between those cells expressing *egfp* and those moving quickly together to the animal pole.

Studying cells speed vectors altogether with their levels of *egfp* expression pointed out that cells expressing *egfp* tend to move faster and collectively. Their

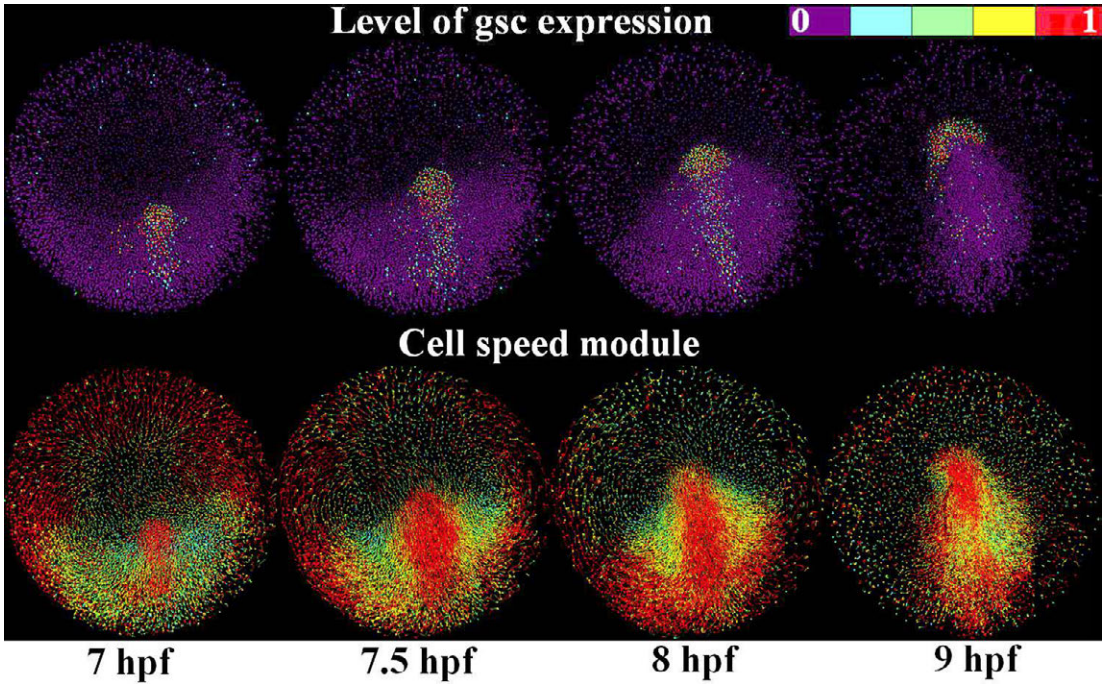

Figure 5.5: *Top row*: *egfp* levels of expression, each colored dot stands up for one cell. *Bottom row*: Speed modules, each colored dot stands up for one current cell whereas the colored lines represent their future positions. Views are taken from underneath the animal pole.

speed vectors steered up coordinate and progressively from bottom to top as the *egfp* expression evolved from the dorsal side to the animal pole. In this sense, we could imagine the *egfp* expression as a "passenger" that walks ahead inside the "train" composed by the migrating cells.

5.4.2 Results: Linking lineage and cell gene expression

Ten clones, chosen to representatively sample the population of migrating cells, were manually validated and labeled, including the six children derived from their mitosis, in order to accurately illustrate the evolution of *egfp* expression and cell speed across the lineage tree. Fig. 5.6 shows the complete cell trajectories of the corrected cells indicating their respective starting and ending points. Some representative results, shown in Fig. 5.7, indicate the fact that the level of gene expression is not constant throughout the cell life. In particular, Fig. 5.7a shows

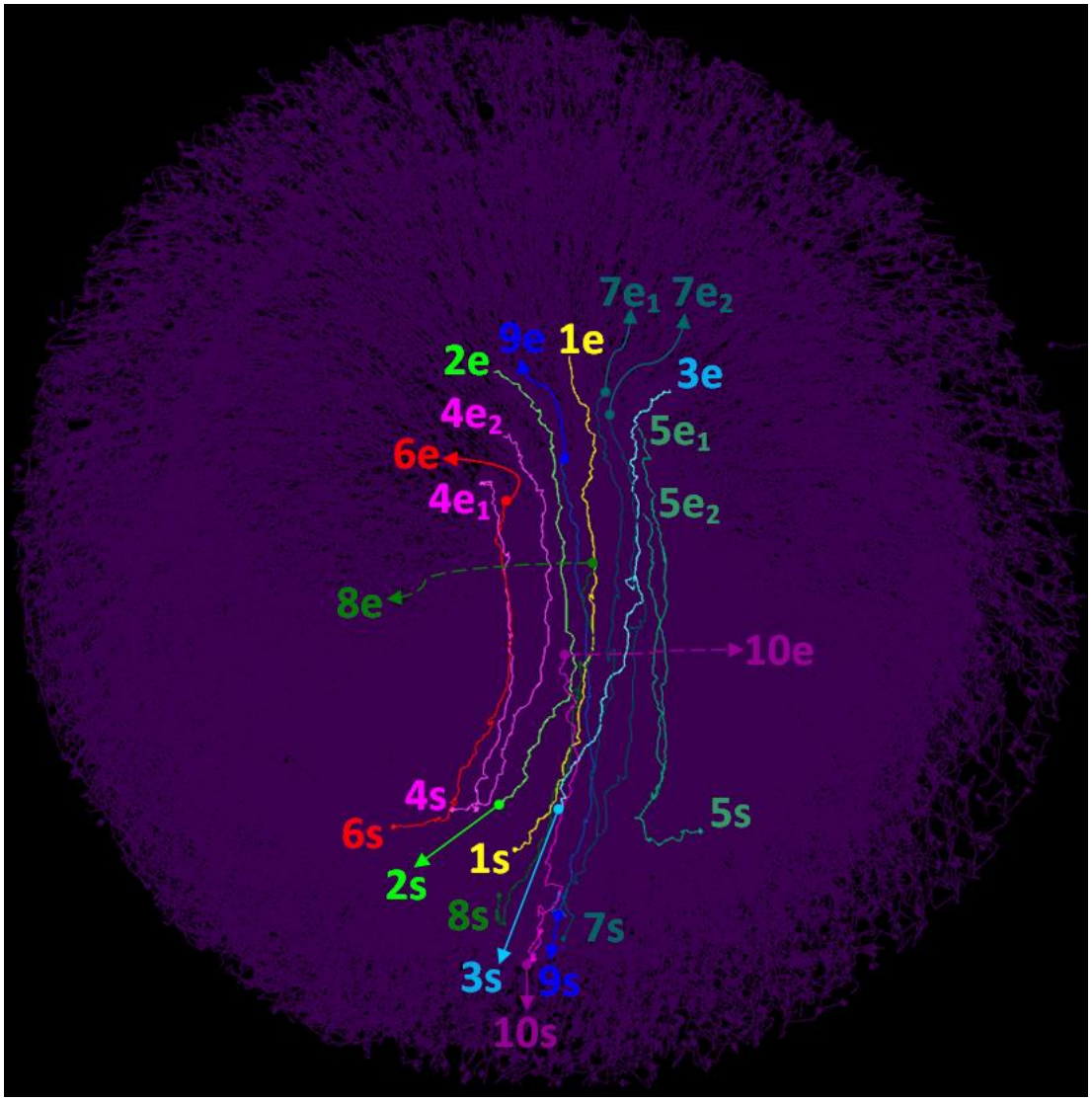

Figure 5.6: Complete trajectories of the 10 validated cells from their start (s) at 6hpf to their end (e) at 9hpf. View taken from underneath the animal pole.

3 cell trajectories with no mitosis: Cell 1, placed at the head of the migrating population shows a strong *egfp* expression from the beginning at 6 hpf to later on decay when the gene has already reached the animal pole at 9 hpf. Cell 9, placed in the middle of the migrating population, only shows *egfp* transcriptional activity during a short period, centered around 7.5 hpf, to rapidly decay afterwards when the *egfp* pattern moves ahead into the animal pole. Cell 10, placed at the tail of

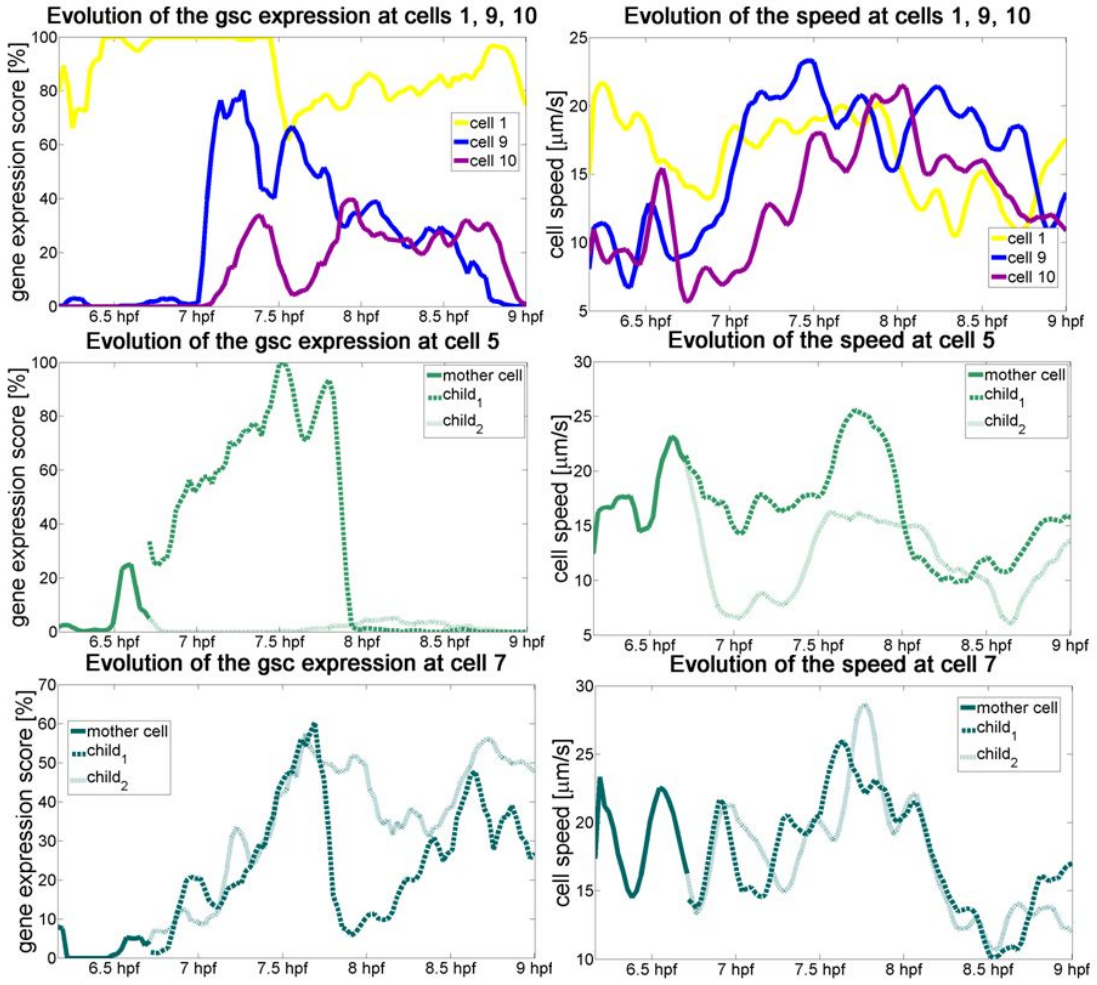

Figure 5.7: Relationship between the levels of *egfp* expression and speed.

the migrating population only starts to timidly show *egfp* activity around 8 hpf when it flocks into the rear part of the *egfp* domain. In the three cases we can see a connection between their levels of *egfp* expression and their speed modules. Fig. 5.7b shows cell 5 undergoing mitosis at 6.7 hpf. Being placed on the margin of the *egfp* domain, one of the children cells starts expressing *egfp* and completes its migration to the animal pole whereas the other do not show any expression and ends up moving slower and finishing its migration earlier. Finally, Fig. 5.7c shows cell 7 undergoing mitosis at 6.7 hpf as well. Being placed on the central area of the *egfp* domain, both children show similar levels of *egfp* expression and follow resembling speeds and trajectories to the animal pole in this case.

These cases suggest that propagation of the *egfp* expression does not follow a clonal behavior but it is rather dynamical since one cell can turn on and off the gene along its life. Therefore, the gene evolution in time can not be adequately mimicked by just propagating its expression through the cell tracking. In other words, *egfp* level of expression is neither constant nor restricted to the same cell population during the embryo development and do not strictly follow a clonal descendance behavior.

5.5 Integration of 3D gene expression atlases into 3D+time digital models

In section 5.4 we describe how to integrate the cell lineage with one gene expression in a live transgenic line. In this section, we introduce a new experiment for combining multiple gene expressions with the lineage tree. In particular, in this experiment we aimed at registering an early 3D atlas of *in situ* gene expression (see section 3.5) into the *in vivo* embryo described in section 5.2 which acts as a live *template* (Fig. 5.9). We opted for this 3D-to-4D registration methodology as it would allow us to profit the wealth of 3D repositories of gene expressions available.

In particular, we focus on the last three developmental stages of the 3D early atlas: 5.3, 6 and 6.3 hpf. Registration of the 3D atlas onto an *in-vivo* dataset (Fig. 5.8) was performed according to methodology presented in section 5.3. In total, 8 different gene expressions, coming from 24 different *in situ* embryos, were matched to the three corresponding developmental stages in the live *template*.

This way, we have the means to study the extent to which gene propagation mechanisms are related to cell descendants (section 5.5.1).

5.5.1 Results: Relating cell lineage to gene propagation and co-expression

Transforming the 3D *in situ* gene expressions onto the *in vivo template* space permits to assess the selection of nuclei considered positive for each expression

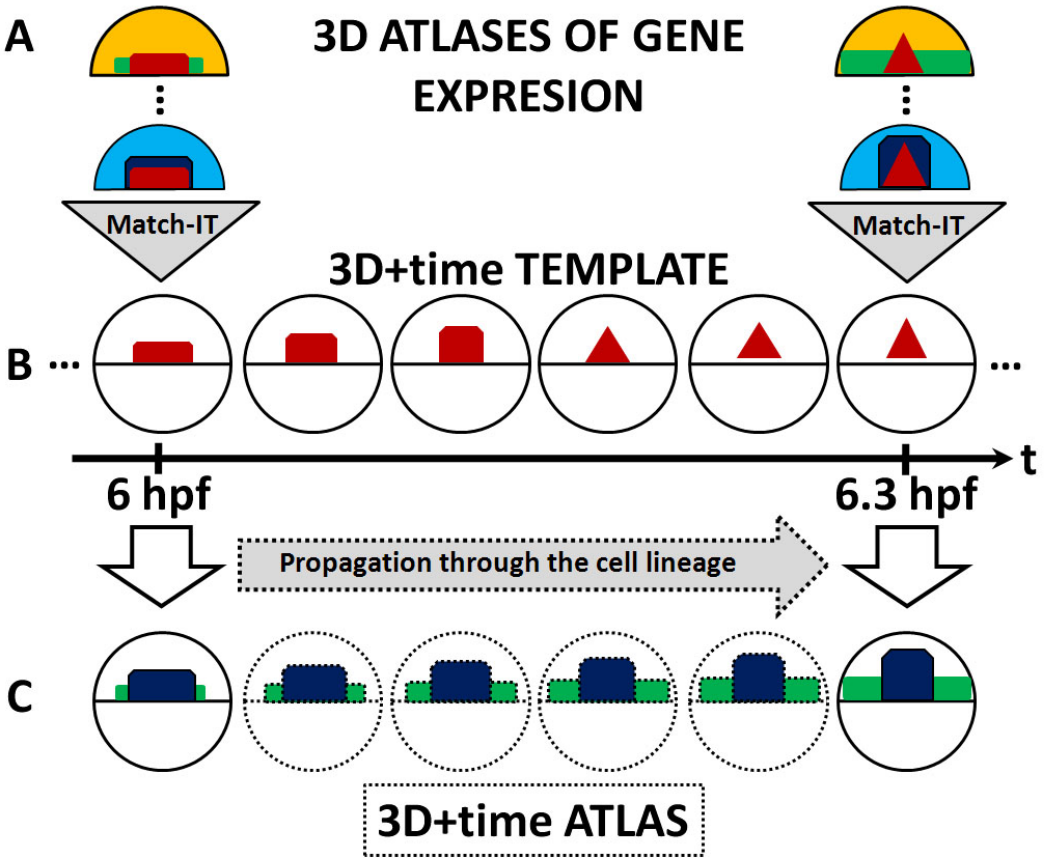

Figure 5.8: Block diagram depicting the integration of a 3D atlas into a 3D+time *template*. The gene expressions labeled at different *analyzed embryos* (**a**), which are acquired *in situ* for two fixed developmental stages (6 and 6.3 hpf), are registered with the Match-IT workflow (see section 3.3) into the *in vivo template* (**b**). (**c**) Depicts the 3D+time propagation of the gene expression domains across the *template* cell lineage.

(section 3.3.4), and then, explore the co-expression between different gene products and visualize how this co-expression evolves in time across the progeny links.

In fact, making a virtual propagation of the 8 *in situ* gene expression patterns at 5.3 hpf through the *template* cell lineage tree, makes it possible to compare how the resulting domains compare to the real gene expression patterns obtained from registering the *in situ* specimens at 6 and 6.3 hpf (Fig. 5.11). In other words, this methodology provides the means to answer the following question: Does gene expression propagate following a "progeny-based" behavior along the cell lineage tree?.

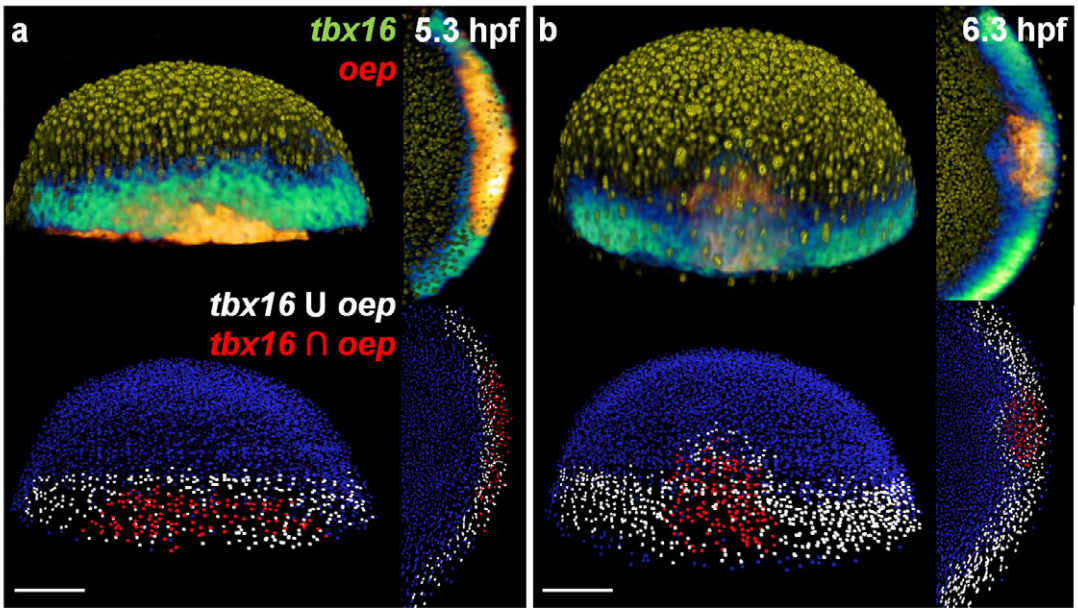

Figure 5.9: Registration of *in situ* gene expressions onto an *in vivo* specimen. *Top row:* Volume rendering of *in situ* gene expressions *oep* (in red) and *tbx16* (in green) superimposed to the *in vivo template* nuclei (in yellow) at 5.3 (**a**) and 6.3 (**b**) hpf. *Bottom row:* Atlas-IT visualization of the nuclei selection positive for both gene expressions (*oep* ∪ *tbx16*, in white) together with the co-expression set (*oep* ∩ *tbx16*, in red). Views are taken from the dorsal side (left) and from the vegetal pole (right). Scale bar 100 μm.

To this end, we selected 44 cells whose trajectories had been automatically tracked entirely from 5.3 to 6.3 hpf (Fig. 5.10). For each of these 44 cells, we can then check, at different times of their development, whether they were considered positive or not for each of the matched *in situ* gene expressions at 5.3, 6 and 6.3 hpf. If gene propagation mechanisms are strictly tied to progeny links, a cell detected to be positive for a gene expression either at 5.3, 6 or 6.3 hpf should also be positive for that expression at the rest of developmental times. However, our first preliminary results indicate that this premise does not hold true for the analyzed genes: most of the 44 cells under study did not keep a stable expression along the lineage, not only in the case of *gsc*, as seen in previous section 5.4, but also for all the other mapped genes, see examples for the *oep*, *ntla* or *tbx16* in Fig. 5.12.

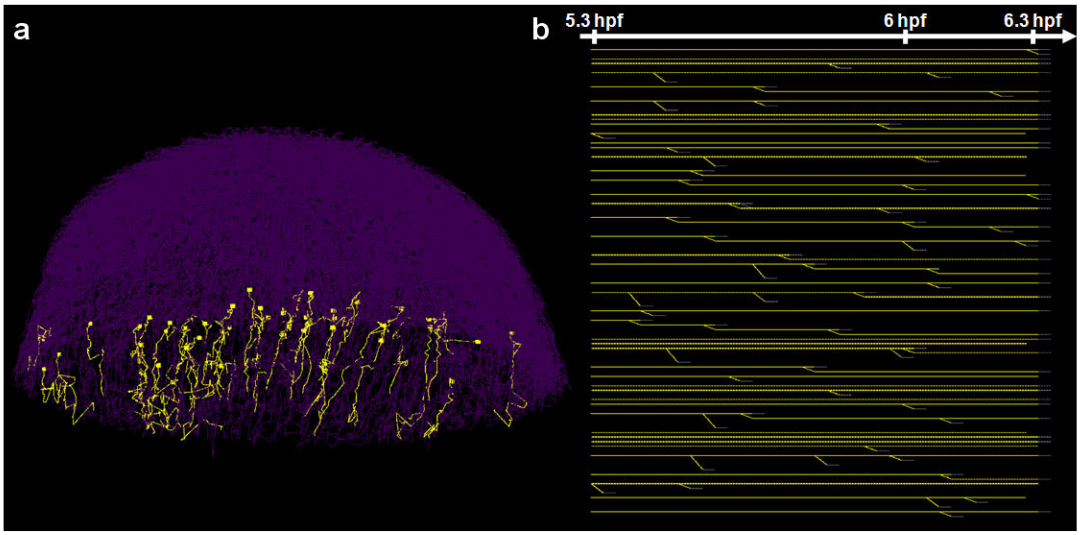

Figure 5.10: (**a**) Representation in Atlas-IT of the spatio-temporal trajectories of 44 cells entirely tracked from 5.3 to 6.3 hpf. During this period cells migrate to the dorsal margin prior to their internalization during the gastrulation process. (**b**) Lineage tree corresponding to these 44 cells.

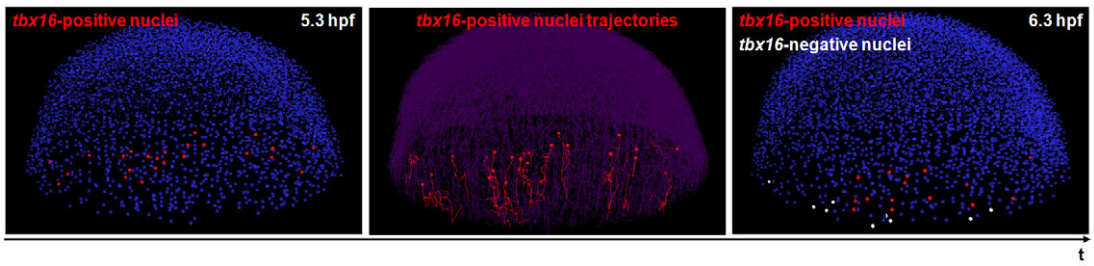

Figure 5.11: *Left panel* shows 24 *template* nuclei (in red) which were found to be positive for *in situ tbx16* expression at 5.3 hpf. *Central panel* shows the propagation of this expression along the selected 24 nuclei trajectories from 5.3 to 6.3 hpf. *Right panel* shows the 15 progeny nuclei at 6.3 hpf which were included in the *in situ tbx16* domain registered at this time point (in red) together with the 9 progeny nuclei which were not determined positive for this *in situ tbx16* expression (in white).

5.6 Conclusions and discussion

We have proposed a computational framework to automatically extract all cells trajectories and their progeny as well as their quantitative levels of expression of *gsc* in a zebrafish embryo undergoing gastrulation between 6 and 9 hpf. We have also registered 8 *in situ* gene expressions to a live *template* specimen at three

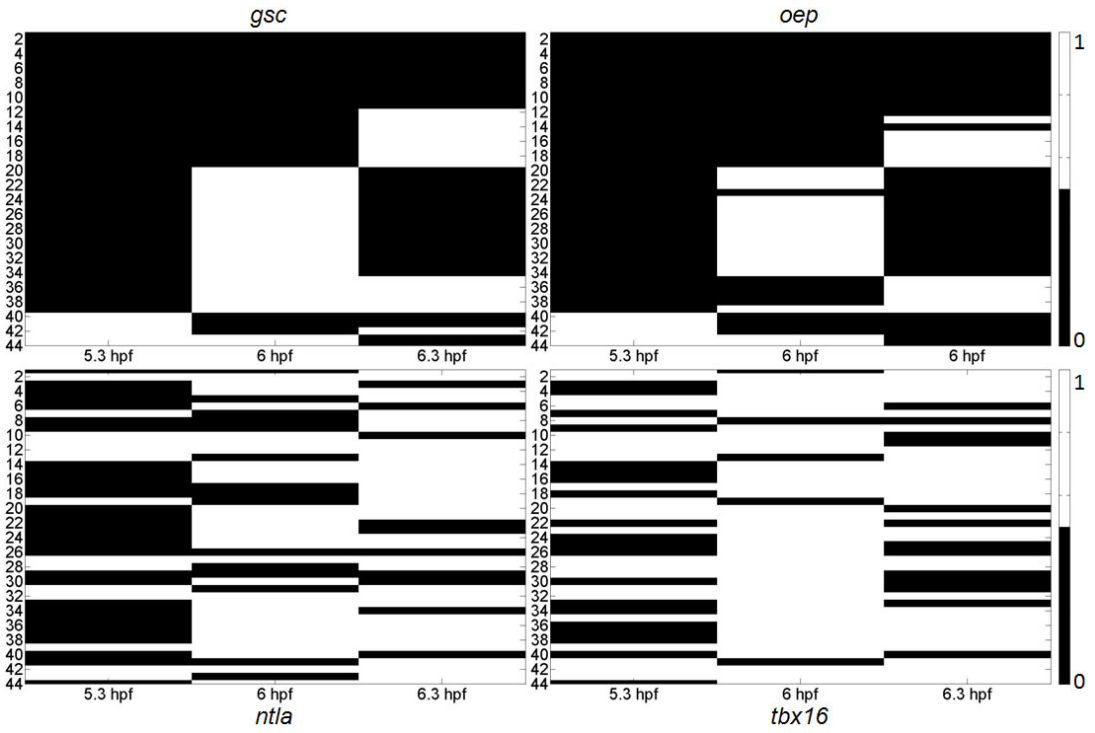

Figure 5.12: Levels of expression of four different gene products -*gsc*, *oep*, *ntla* and *tbx16*- measured for 44 cells at three different developmental stages -5.3, 6 and 6.3 hpf- of their tracked lineage trees.

different developmental stages: 5.3, 6 and 6.3 hpf. This strategy is generic and could be applied to further gene products and developmental stages in an attempt to link gene expression from early (chapter 3) to late (chapter 4) embryogenesis. Similarly, it should also be applicable to public *in situ* gene expression databases such as the Berkeley Drosophila or the Allan Mouse Brain.

The resulting knowledge of the $(x, y, z, t, g_1...g_N)$ space let us establish implications about how genetic expression relates to cell dynamics and descendants. For instance, we obtained hints suggesting that *gsc* gene expression is not deterministically attached to progeny links (section 5.4.2). Indeed, it appeared that gene expression does not exclusively follow cell lineage. Therefore, just propagating gene expression through the tracked itineraries of detected cells does not seem to be an accurate way to model its domain evolution. We also could analyze the correlation between cell migration speed and levels of gene expression.

An interesting future line in this sense would consist in correlating cell populations positive for gene expression with other kinematic and mechanical descriptors apart from the speed (Blanchard et al. [2009], Pastor-Escuredo et al. [2012], Garcia-Ojalvo and Martinez Arias [2012]) that should be extracted from the cell tracking.

Future work includes the development of algorithms that can automatically identify equivalent developmental times between *in situ* embryos and *in vivo templates* by comparison of their corresponding *egfp* labeling.

The proposed strategy mapping a 3D atlas into a 3D+time *template* has a limited scope as it serves to assess whether gene expression propagates following a clonal behavior and to quantify the difference between cell and gene expression dynamics at discrete time points. In order to overcome this limitations, a 3D+time to 3D+time registration methodology should be developed to match various *in vivo* specimens in order to compose a fully spatio-temporal atlas of gene expression. This novel paradigm will have to deal with new methodological and conceptual issues such as the temporal synchronization of embryos developing at different paces and the composition of a statistical, "average" lineage tree. The resulting 3D+time atlas would serve to assess how cell fate is modulated by a particular gene expression profile.

I found myself in the position of that child in a story who noticed a bit of string and -out of curiosity- pulled on it to discover that it was just the tip of a very long and increasingly thick string [...] and kept bringing out wonders beyond reckoning.

Benoît Mandelbrot

Chapter 6

Contributions

6.1 Contributions

The main contributions of this work are listed below:

- New generic computational workflows that allow combining several microscopy images of zebrafish embryos containing gene expression patterns into one single quantitative digital model. Specific atlas reconstruction strategies have been developed for the early and late zebrafish embryogenesis stages, depending on the availability of anatomical references in the specimens and based upon rigid and deformable image registration techniques.

- A new dedicated software platform devoted to visualize and validate the multi-dimensional data of gene expression atlases with cellular resolution. This tool is generic for different animal models and developmental stages. It allows exploring visually raw image data and associated digital segmentations, selections of individual cells positive for a certain gene expression and querying the atlas for co-expression domains. This platform is currently in use at the Institut de Neurobiologie Alfred Fessard (CNRS, France) and has been made freely available to the scientific community.

- A generic set of new analysis algorithms and quantitative measurements devoted to mine digital gene expression atlases. In particular, we have

introduced clustering techniques that automatically group individual cells according to their gene expression profiles and link this information to their anatomical location. Similarly, we have applied clustering techniques to gene expression dynamics (3D+time shapes) in order to extract gene expression families which co-express jointly in space and time. We have also established the new concept of gene expression entropy as a compact way to measure the global gene expression complexity and its evolution through time.

- Two gene expression atlases have been constructed and will be made freely available to the scientific community (release planned during 2013), constituting the first atlas gathering cellular-level gene expression data in early vertebrate embryogenesis and an easily-extendable resource for the study of zefrafish brain development:

    1. A gene atlas of early embryogenesis depicting the spatio-temporal evolution of 8 gene expression domains extracted from 51 individuals during 6 developmental stages evenly distributed between 4.3 and 6.3 *hpf*.

    2. A gene atlas of late brain development depicting the spatial regionalization of 8 gene expression domains extracted from 5 individuals at 48 *hpf*.

- A proof-of-concept framework to explore new strategies towards a 3D+time (4D) gene expression atlas. We have integrated gene expression data and the cell lineage tree by performing the digital reconstruction of a transgenic zebrafish specimen showing both gene and cell nuclei information. In addition, we have demonstrated the feasibility of integrating a 3D gene expression atlas resource into a 3D+time individual acquired *in vivo*. This novel approach opens the possibility of assessing the influence of gene patterning on cellular fate by integrating gene expression data in the cell lineage tree.

6.2 Future work

In the short term, a special focus will be given to the technology transfer of the current research work. As mentioned in previous chapters, a public release of the image-fusion tool (Match-IT), the atlas visualization software (Atlas-IT) and the gene expression cartographies generated is ready to be made. This release has been planned keeping in mind the current trend in the bioimage informatics field towards open-source software and reproducibility (Cardona and Tomancak [2012], Carpenter et al. [2012]). Further integration of these tools within existing processing platforms such as Bioemergences[1], Icy (de Chaumont et al. [2012]) or Fiji (Schindelin et al. [2012]) could also be an obvious expansion of the present work.

Concerning the 3D early embryogenesis atlas, future work includes adapting the processing pipeline for the use of multiplexed *in situ* hybridization acquisitions (Choi et al. [2010]) in order to speed up the inclusion of multiple gene products in the atlas database. The expansion of the current early developmental atlas to later stages, up to 10 hpf, it is also foreseen as well as its application to other embryo stains different from wild type specimens (e.g. *zoep* mutants). This will allow running comparisons on how the gene expressions in wild type and mutant specimens differ in time and space. Additionally, the implementation of a gene quantification scheme (Crombach et al. [2012a], Dubuis et al. [2013]) could overcome the current binary, segmentation-based approach by providing fuzzy, quantitative measurements of gene expression levels. However, these strategies pose serious challenges on how to make robust, comparable quantifications across different embryos, with different staining and imaged under different acquisition conditions. Lastly, the gene expression data contained in the 3D atlas could be used to model both pattern formation under genetic regulation (Doursat [2008]) and the underlying gene regulatory network (GRN) of the zebrafish (Crombach et al. [2012b], Peter et al. [2012]).

Concerning the atlas of the developing zebrafish brain, future work includes the labelization of the *template* brain so that the analysis algorithms introduced in section 3.7 can be applied to mine the resource and make correlations be-

[1] http://bioemergences.iscpif.fr

tween gene expression and brain parcellation. Additionally, the resulting image database is planned to be linked to the interactive visualization interface Atlas-IT. Future developments may also include the implementation of machine-learning techniques to automatically recognize the anatomical landmarks (Stern et al. [2011], Ronneberger et al. [2012]) that guide the initial alignment between different specimens. Given that our methods are labeling-independent, their application to further labeling techniques, such as Brainbow (Livet et al. [2007]), could also be considered in the future.

Finally, concerning the integration of 3D atlases with 3D+time acquisitions, an interesting future line in this area consists in correlating gene expression cellular-level selections with other kinematic (e.g. the speed) and mechanical descriptors (Blanchard et al. [2009], Pastor-Escuredo et al. [2012], Garcia-Ojalvo and Martinez Arias [2012]) that could be extracted from the cell tracking or the optical flow (Amat et al. [2013]).

Additionally, the development of schemes that can automatically recognize corresponding developmental stages between different embryos would be a great step forward to ease registration between *in situ* individuals and *in vivo templates*. This is just an example showing how working with 3D+time atlases will also imply the future use of techniques operating directly in the 4D space (Luengo-Oroz et al. [2012]).

This kind of schemes will contribute to the achievement of the next big challenge in this area: The development of 3D+time to 3D+time registration methods which could directly align different *in vivo* transgenic lines acquired by time-lapse microscopy. This innovative methodology will face major unresolved questions: how to map and integrate spatio-temporal embryos which develop at different paces?, how to compose a statistical lineage tree from the different cell tracking graphs generated at each individual?, how to deal with the great amount of data generated by each live acquisition?. In this sense, it would be interesting to use of an "intelligent" microscope (Conrad et al. [2011]) that, depending on the feedback from automatic algorithms, can adaptably focus on just certain groups of cells or events of interest.

Publications derived from the present PhD Thesis

Journal articles

- C. Castro-González, M.A. Luengo-Oroz, L. Duloquin, T. Savy, B. Rizzi, S. Desnoulez, R. Doursat, Y. Kergosien, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, and A. Santos. "A digital framework to build, visualize and analyze gene expression atlases with single-cell precision in zebrafish early embryogenesis", submitted to *Nature Methods*, March 2013.

- M.A. Luengo-Oroz, D. Pastor-Escuredo, C. Castro-Gonzalez, E. Faure, T. Savy, B. Lombardot, J.L. Rubio-Guivernau, L. Duloquin, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "3D+t Morphological Processing: Applications to Embryogenesis Image Analysis". *IEEE Transactions in Image Processing*, 21(8):3518-3530, August 2012.

- C. Castro-González, M.J. Ledesma-Carbayo, N. Peyriéras, A. Santos. "Assembling Models of Embryo Development: Image Analysis and the Construction of Digital Atlases". *Birth Defects Res. Part C-Embryo Today-Rev.*, 96(2):109-120, June 2012.

- M. Schaap, C.T. Metz, T. Walsum, A.G. Giessen, A.C. Weustink, N.R. Mollet, C. Bauer, H. Bogunovi, C. Castro, X. Deng, E. Dikici, T. O'Donnell, M. Frenay, O. Friman, M. Hernández-Hoyos, P.H. Kitslaar, K. Krissian, C. Kuhnel, M.A. Luengo-Oroz, M. Orkisz, O. Smedby, S. Zambal, Y. Zhang,

G.P. Krestin, W.J. Niessen. "Standardized Evaluation Methodology and Reference Database for Evaluating Coronary Artery Centerline Extraction Algorithms". *Medical Image Analysis*, vol. 13, issue 5, pp. 701-714, October 2009.

Conference proceedings

- C. Castro, M.A. Luengo-Oroz, L. Duloquin, T. Savy, C. Melani, S. Desnoulez, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "Image Processing Challenges in the Creation of Spatiotemporal Gene Expression Atlases of Developing Embryos". *Proc. 33rd Annual International IEEE EMBS Conference*, pp. 5520-5523, Boston, USA, September 2011.

- C. Castro-González, M.A. Luengo-Oroz, L. Duloquin, T. Savy, C. Melani, S. Desnoulez, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "Towards a Digital Model of Zebrafish Embryogenesis. Integration of cell tracking and gene expression quantification". *Proc. 32nd Annual International IEEE EMBS Conference*, pp. 5520-5523, Buenos Aires, Argentine, September 2010.

- C. Castro, M.A. Luengo-Oroz, S. Desnoulez, L. Duloquin, L. Fernández-de-Manuel, S. Montagna, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "An Automatic Quantification and Registration Strategy to Create a Gene Expression Atlas of Zebrafish Embryogenesis". *Proc. 31st Annual International IEEE EMBS Conference*, vol. 1, p. 1469-72, Minneapolis, USA, September 2009.

- M.A. Luengo-Oroz, L. Duloquin, C. Castro, T. Savy, E. Faure, B. Lombardot, P. Bourgine, N. Peyriéras, A. Santos. "Can Voronoi diagram model cell geometries in early sea-urchin embryogenesis?". *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2008)*, pp. 504-507, Paris, France, May 2008.

Other publications

- M.A. Luengo-Oroz, D. Pastor-Escuredo, C. Castro-González, E. Faure, T. Savy, B. Lombardot, J.L. Rubio-Guivernau, L. Duloquin, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "3D+t Morphological Image Analysis for In-vivo Imaging of Embryo Development", *Proc. BioImage Informatics Conference*, pp. 52, Dresden, Germany, September 2012.

- C. Castro, M.A. Luengo-Oroz, L. Duloquin, T. Savy, S. Desnoulez, B. Rizzi, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras and A. Santos. "Registration tools to build a digital 3D cartography of early zebrafish embryogenesis", *9th IEEE International Symposium on Biomedical Imaging (ISBI'12)*, Barcelona, Spain, May 2012.

- C. Castro, M.A. Luengo-Oroz, L. Duloquin, T. Savy, C. Melani, S. Desnoulez, M.J. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, A. Santos. "Linking cell lineage to genetic expression with an automatic quantification and cell tracking framework". *Proc. 2nd Annual International Morphogenesis in Living Systems Conference*, pp. 36, Paris, France, May 2010.

- C. Castro, M.A. Luengo-Oroz, A. Santos, M.J. Ledesma-Carbayo. "Coronary Artery Tracking in 3D Cardiac CT Images Using Local Morphological Reconstruction Operators". *The Insight Journal. 2008 MICCAI Workshop - Grand Challenge Coronary Artery Tracking*, September 2008. http://hdl.handle.net/10380/1436.

References

A. Abbott. Microscopic marvels: Seeing the system. *Nature*, 459:630–631, 2009. 15, 27

M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, et al. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000. 1

P. Aljabar, RA Heckemann, A. Hammers, JV Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009. 9

F. Amat, E.W. Myers, and P.J. Keller. Fast and robust optical flow for time-lapse microscopy using super-voxels. *Bioinformatics*, 29(3):373–380, 2013. 108

J.F. Amatruda, J.L. Shepard, H.M. Stern, and L.I. Zon. Zebrafish as a cancer model system. *Cancer Cell*, 1(3):229–231, 2002. 1

I. Arganda-Carreras, COS Sorzano, P. Thévenaz, A. Muñoz-Barrutia, J. Kybic, R. Marabini, JM Carazo, and C. Ortiz-de Solorzano. Non-rigid consistent registration of 2D image sequences. *Physics in Medicine and Biology*, 55:6215, 2010. 21

X. Artaechevarria, A. Muñoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE Transactions on Medical Imaging*, 28(8):1266–1277, 2009. 2

Albina Asadulina, Aurora Panzera, Csaba Verasztó, Christian Liebig, Gáspár Jékely, et al. Whole-body gene expression pattern registration in Platynereis larvae. *EvoDevo*, 3(1):27, 2012. 21

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. 22

BB Avants, CL Epstein, M. Grossman, and JC Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008. 73, 78

B.B. Avants, N.J. Tustison, G. Song, P.A. Cook, A. Klein, and J.C. Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 73

M. Baker. Screening: the age of fishes. *Nature Methods*, 8(1):47–51, 2010. 2

E. Balanou. Image registration methods for reconstructing a gene expression atlas of early zebrafish embryogenesis. Master's thesis, University of Patras, 2010. 20

R.A. Baldock and A. Burger. Biomedical atlases: Systematics, informatics and analysis. *Advances in Systems Biology*, pages 655–677, 2012. 2

R.A. Baldock, J.B.L. Bard, A. Burger, N. Burton, J. Christiansen, G. Feng, B. Hill, D. Houghton, M. Kaufman, J. Rao, et al. EMAP and EMAGE. *Neuroinformatics*, 1(4):309–325, 2003. 19, 22

Z. Bao, J.I. Murray, T. Boyle, S.L. Ooi, M.J. Sandel, and R.H. Waterston. Automated cell lineage tracing in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 103(8):2707, 2006. 88

TP Barros, WK Alderton, HM Reynolds, AG Roach, and S. Berghmans. Zebrafish: an emerging technology for in vivo pharmacological assessment to identify potential safety liabilities in early drug discovery. *British Journal of Pharmacology*, 154(7):1400–1413, 2008. 1

G.B. Blanchard, A.J. Kabla, N.L. Schultz, L.C. Butler, B. Sanson, N. Gorfinkiel, L. Mahadevan, and R.J. Adams. Tissue tectonics: morphogenetic strain rates,

cell shape change and intercalation. *Nature Methods*, 6(6):458–464, 2009. 104, 108

S. Blanchoud, Y. Budirahardja, F. Naef, and P. Gönczy. ASSET: A robust algorithm for the automated segmentation and standardization of early Caenorhabditis elegans embryos. *Developmental Dynamics*, 239(12):3285–3296, 2010. 20

J. Boline, E.F. Lee, and A.W. Toga. Digital atlases as a framework for data sharing. *Frontiers in Neuroscience*, 2(1):100–106, 2008. 22

T. Boyle, Z. Bao, J. Murray, C. Araya, and R. Waterston. AceTree: a tool for visual analysis of Caenorhabditis elegans embryogenesis. *BMC Bioinformatics*, 7(1):275, 2006. 40

T. Brend and S.A. Holley. Zebrafish whole mount high-resolution double fluorescent in situ hybridization. *Journal of Visualized Experiments: JoVE*, (25): e1229, 2009. doi: 10.3791/1229. 15, 27, 32

S. Bruckner, V. Solteszova, M.E. Groller, J. Hladuvka, K. Buhler, J.Y. Yu, and B.J. Dickson. Braingazer-visual queries for neurobiology research. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1497–1504, 2009. 25

A. Cardona and P. Tomancak. Current challenges in open-source bioimage informatics. *Nature Methods*, 9(7):661–665, 2012. 107

A.E. Carpenter, L. Kamentsky, and K.W. Eliceiri. A call for bioimaging software usability. *Nature Methods*, 9(7):666–670, 2012. 107

J.P. Carson, T. Ju, H.C. Lu, C. Thaller, M. Xu, S.L. Pallas, M.C. Crair, J. Warren, W. Chiu, and G. Eichele. A digital atlas to characterize the mouse brain transcriptome. *PLoS Computational Biology*, 1(4):e41, 2005. 12, 14, 18

C. Castro, MA Luengo-Oroz, S. Desnoulez, L. Duloquin, S. Montagna, MJ Ledesma-Carbayo, P. Bourgine, N. Peyrieras, and A. Santos. An automatic quantification and registration strategy to create a gene expression atlas of zebrafish embryogenesis. In *Proc. IEEE EMBS Conference*, pages 1469–1472, 2009. 10, 13, 16, 20, 40, 88

C. Castro, M.A. Luengo-Oroz, L. Douloquin, T. Savy, C. Melani, S. Desnoulez, P. Bourgine, N. Peyriéras, M.J. Ledesma-Carbayo, and A. Santos. Image processing challenges in the creation of spatiotemporal gene expression atlases of developing embryos. In *Proc. IEEE EMBS Conference*, pages 6841–6844, 2011. 22

C. Castro-González, M.J. Ledesma-Carbayo, N. Peyriéras, and A. Santos. Assembling models of embryo development: Image analysis and the construction of digital atlases. *Birth Defects Research Part C: Embryo Today: Reviews*, 96 (2):109–120, 2012. 11

T.M. Chan, W. Longabaugh, H. Bolouri, H.L. Chen, W.F. Tseng, C.H. Chao, T.H. Jang, Y.I. Lin, S.C. Hung, H.D. Wang, and Yuh C.H. Developmental gene regulatory networks in the zebrafish embryo. *Biochimica et Biophysica Acta (BBA)*, 1789(4):279–298, 2009. 29, 30

A.S. Chiang, C.Y. Lin, C.C. Chuang, H.M. Chang, C.H. Hsieh, C.W. Yeh, C.T. Shih, J.J. Wu, G.T. Wang, Y.C. Chen, et al. Three-Dimensional Reconstruction of Brain-wide Wiring Networks in Drosophila at Single-Cell Resolution. *Current Biology*, 21(1):1–11, 2011. 14

A.T. Chinwalla, L.L. Cook, K.D. Delehaunty, G.A. Fewell, L.A. Fulton, R.S. Fulton, T.A. Graves, L.D.W. Hillier, E.R. Mardis, J.D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915): 520–562, 2002. 1

H.M.T. Choi, J.Y. Chang, L.A. Trinh, J.E. Padilla, S.E. Fraser, and N.A. Pierce. Programmable in situ amplification for multiplexed imaging of mRNA expression,. *Nature Biotechnology*, 28(11):1208–1212, Nov 2010. 2, 27, 60, 62, 107

D.M. Chudakov, S. Lukyanov, and K.A. Lukyanov. Fluorescent proteins as a toolkit for in vivo imaging. *Trends in Biotechnology*, 23(12):605–613, 2005. 15, 27

C. Conrad, A. Wünsche, T.H. Tan, J. Bulkescher, F. Sieckmann, F. Verissimo, A. Edelstein, T. Walter, U. Liebel, R. Pepperkok, et al. Micropilot: automation

of fluorescence microscopy-based imaging for systems biology. *Nature Methods*, 8(3):246–249, 2011. 15, 108

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-interscience, 2006. 56

A. Crombach, D. Cicin-Sain, K.R. Wotton, and J. Jaeger. Medium-throughput processing of whole mount in situ hybridisation experiments into gene expression domains. *PLoS One*, 7(9):e46658, 2012a. 62, 107

A. Crombach, K.R. Wotton, D. Cicin-Sain, M. Ashyraliyev, and J. Jaeger. Efficient reverse-engineering of a developmental gene regulatory network. *PLoS Computational Biology*, 8(7):e1002589, 2012b. 29, 107

S. Damle, B. Hanser, E.H. Davidson, and S.E. Fraser. Confocal quantification of cis-regulatory reporter gene expression in living sea urchin. *Developmental Biology*, 299(2):543–550, 2006. 18

E.H. Davidson and D.H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006. 2

F. de Chaumont, S. Dallongeville, N. Chenouard, N. Hervé, S. Pop, T. Provoost, V. Meas-Yedid, P. Pankajakshan, T. Lecomte, Y. Le Montagner, et al. Icy: an open bioimage informatics platform for extended reproducible research. *Nature Methods*, 9(7):690–696, 2012. 107

Doursat R. & Peyriéras N. Delile, J. *From cell behavior to tissue deformation: Computational modeling and simulation of early animal embryogenesis with the MecaGen platform*. Computational Systems Biology, 2nd edition, Academic Press, 2013. 29

Piotr Dollár. Matlab TPS (Thin Plate Spline) implementation. `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`, 2006. 4

AE Dorr, JP Lerch, S. Spring, N. Kabani, and RM Henkelman. High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult C57Bl/6J mice. *Neuroimage*, 42(1):60–69, 2008. 18, 19

R. Doursat. The self-made puzzle: Integrating self-assembly and pattern formation under non-random genetic regulation. *InterJournal Complex Systems*, (2292), 2008. 107

O. Drblikova, M. Komornikova, M. Remesikova, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarte. Estimate of the cell number growth rate using PDE methods of image processing and time series analysis. *Journal of Electrical Engineering*, 58(7):86–92, 2007. 19, 91

J.O. Dubuis, R. Samanta, and T. Gregor. Accurate measurements of dynamics and reproducibility in small genetic networks. *Molecular Systems Biology*, 9: 639, 2013. doi: 10.1038/msb.2012.72. 62, 107

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998. 52, 61

C. Elegans Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012–2018, 1998. 1

K.W. Eliceiri, M.R. Berthold, I.G. Goldberg, L. Ibáñez, BS Manjunath, M.E. Martone, R.F. Murphy, H. Peng, A.L. Plant, B. Roysam, et al. Biological imaging software tools. *Nature Methods*, 9(7):697–710, 2012. 25

Project Ensembl. Ensembl genome browser 59: Danio rerio. `http://www.ensembl.org/Danio_rerio/Info/Index/`, 2007. 1

D. Evanko. Microscope harmonies. *Nature Methods*, 7(10):779, 2010. 15

A.C. Evans, A.L. Janke, L.D. Collins, and S. Baillet. Brain templates and atlases. *NeuroImage*, 62(2):911–922, 2012. 66

R. Fernandez-Gonzalez, A. Munoz-Barrutia, M.H. Barcellos-Hoff, and C. Ortiz-de Solorzano. Quantitative in vivo microscopy: the return from the 'omics'. *Current Opinion in Biotechnology*, 17(5):501–510, 2006. 2

M. Fisher, H. Downie, M.C.M. Welten, I. Delgado, A. Bain, T. Planzer, A. Sherman, H. Sang, and C. Tickle. Comparative analysis of 3D expression patterns

of transcription factor genes and digit fate maps in the developing chick wing. *PLoS One*, 6(4):e18661, 2011. 10, 13, 14, 24

M.E. Fisher, A.K. Clelland, A. Bain, R.A. Baldock, P. Murphy, H. Downie, C. Tickle, D.R. Davidson, and R.A. Buckland. Integrating technologies for comparing 3D gene expression domains in the developing chick limb. *Developmental Biology*, 317(1):13–23, 2008. 10, 16, 19

C.G. Fonseca, M. Backhaus, D.A. Bluemke, R.D. Britten, J. Do Chung, B.R. Cowan, I.D. Dinov, J.P. Finn, P.J. Hunter, A.H. Kadish, et al. The cardiac atlas project - an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics*, 27(16):2288–2295, 2011. 9

C.C. Fowlkes, C.L.L. Hendriks, S.V.E. Keranen, G.H. Weber, O. Rubel, M.Y. Huang, S. Chatoor, A.H. DePace, L. Simirenko, et al. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, 133 (2):364–374, 2008. 10, 13, 18, 22, 24, 28, 39, 87

C.C. Fowlkes, K.B. Eckenrode, M.D. Bragdon, M. Meyer, Z. Wunderlich, L. Simirenko, C.L.L. Hendriks, S.V.E. Keränen, C. Henriquez, D.W. Knowles, et al. A conserved developmental patterning network produces quantitatively different output in multiple species of drosophila. *PLoS Genetics*, 7(10): e1002346, 2011. 29

E. Frise, A.S. Hammonds, and S.E. Celniker. Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. *Molecular Systems Biology*, 6:345, 2010. 10, 14, 16, 18, 19, 62

P. Frolkovic, K. Mikula, N. Peyriéras, and A. Sarti. Counting number of cells and cell segmentation using advection-diffusion equations. *KYBERNETIKA*, 43(6):817–829, 2007. 19

J. Garcia-Ojalvo and A. Martinez Arias. Towards a statistical mechanics of cell fate decisions. *Current Opinion in Genetics & Development*, 22(6):619–626, 2012. 104, 108

B.N.G. Giepmans, S.R. Adams, M.H. Ellisman, and R.Y. Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312(5771):217–224, 2006. 15

P. Goldsmith. Zebrafish as a pharmacological tool: the how, why and when. *Current Opinion in Pharmacology*, 4(5):504–512, 2004. 1

N. Gorfinkiel, G.B. Blanchard, R.J. Adams, and A.M. Arias. Mechanical control of global cell behaviour during dorsal closure in Drosophila. *Development*, 136 (11):1889–1898, 2009. 3

N. Gorfinkiel, S. Schamberg, and G.B. Blanchard. Integrative approaches to morphogenesis: lessons from dorsal closure. *Genesis*, 49(7):522–533, 2011. 2

A. Gouaillard, T. Brown, M. Bronner-Fraser, S.E. Fraser, and S.G. Megason. GoFigure and The Digital Fish Project: Open tools and open data for an imaging based approach to system biology. *Insight Journal. Available at: http://hdl. handle. net/1926/565*, 2007. 25, 40

M. Hawrylycz, R.A. Baldock, A. Burger, T. Hashikawa, G.A. Johnson, M. Martone, L. Ng, C. Lau, S.D. Larsen, J. Nissanov, et al. Digital atlasing and standardization in the mouse brain. *PLoS Computational Biology*, 7(2):e1001065, 2011. 12

M.J. Hawrylycz, S. Lein, A.L. Guillozet-Bongaarts, E.H. Shen, L. Ng, J.A. Miller, L.N. van de Lagemaat, K.A. Smith, A. Ebbert, Z.L. Riley, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489 (7416):391–399, 2012. 12

S.B. Hedges, J. Dudley, and S. Kumar. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, 2006. 12

M. Held, M.H.A. Schmitz, B. Fischer, T. Walter, B. Neumann, M.H. Olma, M. Peter, J. Ellenberg, and D.W. Gerlich. Cellcognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nature Methods*, 7(9): 747–754, 2010. 40

C.L.L. Hendriks, S.V.E. Keränen, C.C. Fowlkes, L. Simirenko, G.H. Weber, A.H. DePace, C. Henriquez, D.W. Kaszuba, B. Hamann, M.B. Eisen, et al. Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biology*, 7(12):R124, 2006. 18, 28

A.J. Hill, H. Teraoka, W. Heideman, and R.E. Peterson. Zebrafish as a model vertebrate for investigating chemical toxicity. *Toxicological Sciences*, 86(1): 6–19, 2005. 10

J. Huisken and D.Y.R. Stainier. Selective plane illumination microscopy techniques in developmental biology. *Development*, 136(12):1963–1975, 2009. 15

R.E. Jacobs, C. Papan, S. Ruffins, J.M. Tyszka, and S.E. Fraser. MRI: volumetric imaging for vital imaging and atlas construction. *Nature Reviews Molecular Cell Biology*, 4:SS10–SS16, 2003. 15

J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, et al. Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430(6997):368–371, 2004. 29

G.A. Johnson, A. Badea, J. Brandenburg, G. Cofer, B. Fubara, S. Liu, and J. Nissanov. Waxholm Space: An image-based reference for coordinating mouse brain research. *Neuroimage*, 53(2):365–372, 2010. 12, 18, 19, 21

T. Jones, I. Kang, D. Wheeler, R. Lindquist, A. Papallo, D. Sabatini, P. Golland, and A. Carpenter. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics*, 9(1):482, 2008. doi: 10.1186/1471-2105-9-482. 25, 40

P.J. Keller and E.H.K. Stelzer. Quantitative in vivo imaging of entire embryos with digital scanned laser light sheet fluorescence microscopy. *Current Opinion in Neurobiology*, 18(6):624–632, 2008. 5, 37

P.J. Keller, A.D. Schmidt, J. Wittbrodt, and E.H.K. Stelzer. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science*, 322(5904):1065, 2008. 88

P.J. Keller, A.D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E.H.K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nature Methods*, 7(8):637–642, 2010. 15

J. Kerwin, Y. Yang, P. Merchan, S. Sarma, J. Thompson, X. Wang, J. Sandoval, L. Puelles, R. Baldock, and S. Lindsay. The HUDSEN atlas: a three-dimensional (3D) spatial framework for studying gene expression in the developing human brain. *Journal of Anatomy*, 217(4):289–299, 2010. 12

K. Khairy and P.J. Keller. Reconstructing embryonic development. *Genesis*, 49 (7):488–513, 2011. 22, 28

C.B. Kimmel, W.W. Ballard, S.R. Kimmel, B. Ullmann, and T.F. Schilling. Stages of embryonic development of the zebrafish. *American Journal of Anatomy*, 203(3):253–310, 1995. 31, 32, 37, 43, 67, 88

A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M.C. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802, 2009. 4, 73

N. Kovačević, JT Henderson, E. Chan, N. Lifshitz, J. Bishop, AC Evans, RM Henkelman, and XJ Chen. A three-dimensional MRI atlas of the mouse brain with estimates of the average and variability. *Cerebral Cortex*, 15(5): 639–645, 2005. 12, 14, 18, 19, 21

Z. Krivá, K. Mikula, N. Peyriéras, B. Rizzi, A. Sarti, and O. Stašová. 3D early embryogenesis image filtering by nonlinear partial differential equations. *Medical Image Analysis*, 14(4):510–526, 2010. 19

Tetsuhiro Kudoh, Michael Tsang, Neil A Hukriede, Xiongfong Chen, Michael Dedekian, Christopher J Clarke, Anne Kiang, Stephanie Schultz, Jonathan A Epstein, Reiko Toyama, et al. A gene expression screen in zebrafish embryogenesis. *Genome Research*, 11(12):1979–1987, 2001. 56

C. Lau, L. Ng, C. Thompson, S. Pathak, L. Kuan, A. Jones, and M. Hawrylycz. Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, 9(1):153, 2008. 25

E. Lécuyer and P. Tomancak. Mapping the gene expression universe. *Current Opinion in Genetics & Development*, 18(6):506–512, 2008. 2

E.S. Lein, M.J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A.F. Boe, M.S. Boguski, K.S. Brockway, E.J. Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007. 12, 13, 14, 20, 21, 28, 62, 87

E. Li and E.H. Davidson. Building developmental gene regulatory networks. *Birth Defects Research Part C: Embryo Today: Reviews*, 87(2):123–130, 2009. 2

X. Liu, F. Long, H. Peng, S.J. Aerni, M. Jiang, A. Sánchez-Blanco, J.I. Murray, E. Preston, B. Mericle, S. Batzoglou, et al. Analysis of cell fate from single-cell gene expression profiles in C. elegans. *Cell*, 139(3):623–633, 2009. 10, 14, 18, 21, 62

J. Livet, T.A. Weissman, H. Kang, J. Lu, R.A. Bennis, J.R. Sanes, and J.W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, 2007. 108

B. Lombardot, M.A. Luengo-Oroz, C. Melani, E. Faure, A. Santos, N. Peyrieras, M. Ledesma-Carbayo, and P. Bourgine. Evaluation of four 3D non rigid registration methods applied to early zebrafish development sequences. In *Microscopic Image Analysis with Applications in Biology, MICCAI*, 2008. 76

F. Long, H. Peng, X. Liu, S.K. Kim, and E. Myers. A 3D digital atlas of C. elegans and its application to single-cell analyses. *Nature Methods*, 6(9):667–672, 2009. 10, 13, 16, 19, 24, 28, 87

F. Long, J. Zhou, and H. Peng. Visualization and Analysis of 3D Microscopic Images. *PLoS Computational Biology*, 8(6):e1002519, 2012. 24

W.J.R. Longabaugh, E.H. Davidson, and H. Bolouri. Visualization, documentation, analysis, and communication of large scale gene regulatory networks. *Biochimica et Biophysica Acta (BBA)*, 1789(4):363, 2009. 29

J.W. Lu, Y. Hsia, H.C. Tu, Y.C. Hsiao, W.Y. Yang, H.D. Wang, and C.H. Yuh. Liver development and cancer formation in zebrafish. *Birth Defects Research Part C: Embryo Today: Reviews*, 93(2):157–172, 2011. 1

M. Luengo-Oroz, D. Pastor-Escuredo, C. Castro-González, E. Faure, T. Savy, B. Lombardot, J. Rubio-Guivernau, L. Duloquin, M. Ledesma-Carbayo, P. Bourgine, N. Peyriéras, and A. Santos. 3D+t morphological processing: Applications to embryogenesis image analysis. *IEEE Transactions on Image Processing*, 21(8):3518–3530, 2012. 108

M.A. Luengo-Oroz. *Mathematical methods for processing and analyzing in-vivo fluorescence images of embryo development*. PhD thesis, Universidad Politécnica de Madrid, 2009. 2

MA Luengo-Oroz, L. Duloquin, C. Castro, T. Savy, E. Faure, B. Lombardot, R. Bourgine, N. Peyriéras, and A. Santos. Can voronoi diagram model cell geometries in early sea-urchin embryogenesis? In *Proc. IEEE ISBI Conference*, pages 504–507, 2008. 19, 92

MA Luengo-Oroz, MJ Ledesma-Carbayo, N. Peyriéras, and A. Santos. Image analysis for understanding embryo development: a bridge from microscopy to biological insights. *Current Opinion in Genetics & Development*, 21(5):630–637, 2011. 13, 28

Y. Ma, PR Hof, SC Grant, SJ Blackband, R. Bennett, L. Slatest, MD McGuigan, and H. Benveniste. A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Neuroscience*, 135 (4):1203–1215, 2005. 12, 18, 19

D.L. Mace, J.Y. Lee, R.W. Twigg, J. Colinas, P.N. Benfey, and U. Ohler. Quantification of transcription factor expression from Arabidopsis images. *Bioinformatics*, 22(14):e323–e331, 2006. 22

A. MacKenzie-Graham, E.F. Lee, I.D. Dinov, M. Bota, D.W. Shattuck, S. Ruffins, H. Yuan, F. Konstantinidis, A. Pitiot, Y. Ding, et al. A multimodal, multidimensional atlas of the C57BL/6J mouse brain. *Journal of Anatomy*, 204(2): 93–102, 2004. 12, 21

P. Mahou, M. Zimmerley, K. Loulier, K.S. Matho, G. Labroille, X. Morin, W. Supatto, J. Livet, D. Debarre, and E. Beaurepaire. Multicolor two-photon tissue imaging by wavelength mixing. *Nature Methods*, 9:815–818, 2012. 15, 60, 62

JB Maintz and M.A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998. 20, 28

N. Malpica, C. Ortiz de Solorzano, J.J. Vaquero, A. Santos, I. Vallcorba, J.M. Garcia-Sagredo, and F. del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289–297, 1997. 19

M.E. Martone, I. Zaslavsky, A. Gupta, A. Memon, J. Tran, W. Wong, L. Fong, S.D. Larson, and M.H. Ellisman. The smart atlas: spatial and semantic strategies for multiscale integration of brain data. *Anatomy Ontologies for Bioinformatics*, pages 267–286, 2008. 1, 2

J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001. 9

A. McMahon, W. Supatto, S.E. Fraser, and A. Stathopoulos. Dynamic analyses of Drosophila gastrulation provide insights into collective cell migration. *Science*, 322(5907):1546–1150, 2008. 14

J.D. McPherson, M. Marra, L.D. Hillier, R.H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E.R. Mardis, R.K. Wilson, et al. A physical map of the human genome. *Nature*, 409(6822):934–941, 2001. 1

S.G. Megason and S.E. Fraser. Digitizing life at the level of the cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development. *Mechanisms of Development*, 120(11):1407–1420, 2003. 88

S.G. Megason and S.E. Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007. 2, 27, 87

C. Melani, M. Campana, B. Lombardot, B. Rizzi, F. Veronesi, C. Zanella, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarti. Cells tracking in a live zebrafish embryo. In *Proc. IEEE EMBS Conference*, pages 1631–1634, 2007. 91

K. Mikula, N. Peyriéras, M. Remesíková, and O. Stasová. Segmentation of 3D cell membrane images by PDE methods and its applications. *Computers in Biology and Medicine*, 41(6):326–339, 2011. 19

D.J. Milan, T.A. Peterson, J.N. Ruskin, R.T. Peterson, and C.A. MacRae. Drugs that induce repolarization abnormalities cause bradycardia in zebrafish. *Circulation*, 107(10):1355–1358, 2003. 1

K.R. Mosaliganti, R.R. Noche, F. Xiong, I.A. Swinburne, and S.G. Megason. ACME: Automated Cell Morphology Extractor for Comprehensive Reconstruction of Cell Membranes. *PLoS Computational Biology*, 8(12):e1002780, 2012. 19

J.I. Murray, Z. Bao, T.J. Boyle, M.E. Boeck, B.L. Mericle, T.J. Nicholas, Z. Zhao, M.J. Sandel, and R.H. Waterston. Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nature Methods*, 5(8):703–709, 2008. 5, 10, 13, 18

G. Myers. Why bioimage informatics matters. *Nature Methods*, 9(7):659–660, 2012. 2

L. Ng, S.D. Pathak, C. Kuan, C. Lau, H. Dong, A. Sodt, C. Dang, B. Avants, P. Yushkevich, J.C. Gee, et al. Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):382–393, 2007. 21

L. Ng, A. Bernard, C. Lau, C.C. Overly, H.W. Dong, C. Kuan, S. Pathak, S.M. Sunkin, C. Dang, J.W. Bohland, et al. An anatomic gene expression atlas of the adult mouse brain. *Nature Neuroscience*, 12(3):356–362, 2009. 21

C. Niehrs and N. Pollet. Synexpression groups in eukaryotes. *Nature*, 402(6761): 483–487, 1999. 47

T.E. North, W. Goessling, C.R. Walkley, C. Lengerke, K.R. Kopani, A.M. Lord, G.J. Weber, T.V. Bowman, I.H. Jang, T. Grosser, et al. Prostaglandin e2 regulates vertebrate haematopoietic stem cell homeostasis. *Nature*, 447(7147): 1007–1011, 2007. 1

A.C. Oates, N. Gorfinkiel, M. González-Gaitán, and C.P. Heisenberg. Quantitative approaches in developmental biology. *Nature Reviews Genetics*, 10(8): 517–530, 2009. 2, 27

N. Olivier, M.A. Luengo-Oroz, L. Duloquin, E. Faure, T. Savy, I. Veilleux, X. Solinas, D. Débarre, P. Bourgine, A. Santos, et al. Cell lineage reconstruction of early zebrafish embryos using label-free nonlinear microscopy. *Science*, 329 (5994):967–971, 2010. 5, 25, 28, 36, 88

D. Paquet, R. Bhat, A. Sydow, E.M. Mandelkow, S. Berg, S. Hellberg, J. Fälting, M. Distel, R.W. Köster, B. Schmid, et al. A zebrafish model of tauopathy allows in vivo imaging of neuronal cell death and drug evaluation. *The Journal of Clinical Investigation*, 119(5):1382, 2009. 13

C. Pardo-Martin, T.Y. Chang, B.K. Koo, C.L. Gilleland, S.C. Wasserman, and M.F. Yanik. High-throughput in vivo vertebrate screening. *Nature Methods*, 7 (8):634–636, 2010. 15

H. Park, P.H. Bland, and C.R. Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *IEEE Transactions on Medical Imaging*, 22(4):483–492, 2003. 9

D. Pastor, MA Luengo-Oroz, B. Lombardot, I. Gonzalvez, L. Duloquin, T. Savy, P. Bourgine, N. Peyrieras, and A. Santos. Cell tracking in fluorescence images of embryogenesis processes with morphological reconstruction by 4D-tubular

structuring elements. In *Proc. IEEE EMBS Conference*, pages 970–973, 2009. 14

D. Pastor-Escuredo, T. Savy, J.C. del Alamo, N. Peyrieras, M.J. Ledesma-Carbayo, and A. Santos. In-vivo multi-scale analysis of embryogenesis kinematics. In *IEEE ISBI Conference*, 2012. 104, 108

B.Y. Paul, C.C. Hong, C. Sachidanandan, J.L. Babitt, D.Y. Deng, S.A. Hoyng, H.Y. Lin, K.D. Bloch, and R.T. Peterson. Dorsomorphin inhibits BMP signals required for embryogenesis and iron metabolism. *Nature Chemical Biology*, 4 (1):33–41, 2008. 1

H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827, 2008. 2

H. Peng, F. Long, X. Liu, S.K. Kim, and E.W. Myers. Straightening caenorhabditis elegans images. *Bioinformatics*, 24(2):234–242, 2008. 20

H. Peng, Z. Ruan, F. Long, J.H. Simpson, and E.W. Myers. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nature Biotechnology*, 28(4):348–353, 2010. 25, 40

H. Peng, P. Chung, F. Long, L. Qu, A. Jenett, A.M. Seeds, E.W. Myers, and J.H. Simpson. BrainAligner: 3D registration atlases of Drosophila brains. *Nature Methods*, 8(6):493–498, 2011. 4, 10, 13, 18, 20, 21, 24, 28, 29, 39, 66, 73, 85

I.S. Peter, E. Faure, and E.H. Davidson. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences*, 109(41):16434–16442, 2012. 2, 29, 107

A. Pisarev, E. Poustelnikova, M. Samsonova, and J. Reinitz. Flyex, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research*, 37(suppl 1):D560–D566, 2009. 25

S. Pop, A. Dufour, J.F. Le Garrec, C.V. Ragni, C. Cimper, S.M. Meilhac, and J.C. Olivo-Marin. Extracting 3D cell parameters from dense tissue environments: Application to the development of the mouse heart. *Bioinformatics*, 2013. doi: 10.1093/bioinformatics/btt027. 85

D. Potikanond and FJ Verbeek. 3D Visual Integration of Spatio-Temporal Gene Expression Patterns on Digital Atlas of Zebrafish Embryo using Web Service. In *Proc. Int. Conf. Advances in Communication and Information Technology*, pages 56–62, 2011. 10, 22

L. Qu and H. Peng. A principal skeleton algorithm for standardizing confocal images of fruit fly nervous systems. *Bioinformatics*, 26(8):1091–1097, 2010. 20

M. Rath, R. Nitschke, A. Filippi, O. Ronneberger, and W. Driever. Generation of high quality multi-view confocal 3D datasets of zebrafish larval brains suitable for analysis using Virtual Brain Explorer (ViBE-Z) software. *Nature Protocol Exchange*, (22/06/2012), 2012. doi: doi:10.1038/protex.2012.031. 59, 66

D.E. Rex, J.Q. Ma, and A.W. Toga. The loni pipeline processing environment. *Neuroimage*, 19(3):1033–1048, 2003. 12, 16, 18, 20

L. Richardson, S. Venkataraman, P. Stevenson, Y. Yang, N. Burton, J. Rao, M. Fisher, R.A. Baldock, D.R. Davidson, and J.H. Christiansen. Emage mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Research*, 38(suppl 1):D703–D709, 2010. 12, 16

J. Rittscher, A. Yekta, O.B. Musodiq, J. Tu, and L.S. Wen. Automated systems and methods for screening zebrafish. US Patent No. 2010/0119119 A1, May 2010. 21

J. Rittscher, D. Padfield, A. Santamaria, J. Tu, A. Can, M. Bello, D. Gao, A. Sood, M. Gerdes, and F. Ginty. Methods and algorithms for extracting high-content signatures from cells, tissues, and model organisms. In *Proc. IEEE ISBI Conference*, pages 1712–1716, 2011. 10, 19

O. Ronneberger, K. Liu, M. Rath, D. Rueβ, T. Mueller, H. Skibbe, B. Drayer, T. Schmidt, A. Filippi, R. Nitschke, et al. ViBE-Z: a framework for 3D virtual colocalization analysis in zebrafish larval brains. *Nature Methods*, 9(7):735–742, 2012. 10, 13, 21, 29, 59, 66, 73, 82, 84, 85, 86, 88, 108

O. Rübel, G.H. Weber, M.Y. Huang, E.W. Bethel, M.D. Biggin, C.C. Fowlkes, C.L. Luengo Hendriks, S.V.E. Keranen, M.B. Eisen, D.W. Knowles, et al. Inte-

grating data clustering and visualization for the analysis of 3D gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):64–79, 2010. 25, 40

J.L. Rubio-Guivernau, V. Gurchenkov, M.A. Luengo-Oroz, L. Duloquin, P. Bourgine, A. Santos, N. Peyrieras, and M.J. Ledesma-Carbayo. Wavelet-based image fusion in multi-view three-dimensional microscopy. *Bioinformatics*, 28(2):238–245, 2012. 23, 75

Seth W Ruffins, Melanie Martin, Lindsey Keough, Salina Truong, Scott E Fraser, Russell E Jacobs, and Rusty Lansford. Digital three-dimensional atlas of quail development using high-resolution MRI. *TheScientificWorldJOURNAL*, 7:592–604, 2007. 10, 16, 19

S. Saalfeld, R. Fetter, A. Cardona, and P. Tomancak. Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nature Methods*, 9(7): 717–720, 2012. 21

T. Savy and Bioemergences. Mov-IT: Morphogenesis Visualization Tool. Unpublished, 2013. 85, 91, 94

A.F. Schier and W.S. Talbot. Molecular genetics of axis formation in zebrafish. *Annu. Rev. Genet.*, 39:561–613, 2005. 27, 42

J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012. 107

J. Serra. Image analysis and mathematical morphology. *Academic Press, London*, I,II., 1982,1988. 37

J. Sharpe, U. Ahlgren, P. Perry, B. Hill, A. Ross, J. Hecksher-Sørensen, R. Baldock, and D. Davidson. Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science*, 296(5567):541–545, 2002. 15

N.S. Sipes, S. Padilla, and T.B. Knudsen. Zebrafish as an integrative model for twenty-first century toxicity testing. *Birth Defects Research Part C: Embryo Today: Reviews*, 93(3):256–267, 2011. 1

S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004. 12, 20, 21

E. Sodergren, G.M. Weinstock, E.H. Davidson, R.A. Cameron, R.A. Gibbs, R.C. Angerer, L.M. Angerer, M.I. Arnone, D.R. Burgess, R.D. Burke, et al. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, 314(5801): 941–52, 2006. 1

C. Sorzano, C. Messaoudi, M. Eibauer, J.R. Bilbao-Castro, R. Hegerl, S. Nickell, S. Marco, and J.M. Carazo. Marker-free image registration of electron tomography tilt-series. *BMC Bioinformatics*, 10(1):124, 2009. 23

H.M. Stern and L.I. Zon. Cancer genetics and drug discovery in the zebrafish. *Nature Reviews Cancer*, 3(7):533–539, 2003. 1, 29

O. Stern, R. Marée, J. Aceto, N. Jeanray, M. Muller, L. Wehenkel, and P. Geurts. Automatic localization of interest points in zebrafish images with tree-based methods. *Pattern Recognition in Bioinformatics*, pages 179–190, 2011. 86, 108

W. Supatto, A. McMahon, S.E. Fraser, and A. Stathopoulos. Quantitative imaging of collective cell migration during drosophila gastrulation: multiphoton microscopy and computational analysis. *Nature Protocols*, 4(10):1397–1412, 2009. 14

W. Supatto, T.V. Truong, D. Débarre, and E. Beaurepaire. Advances in multiphoton microscopy for imaging embryos. *Current Opinion in Genetics & Development*, pages 538–548, 2011. 27

A.W. Toga, P.M. Thompson, S. Mori, K. Amunts, and K. Zilles. Towards multimodal atlases of the human brain. *Nature Reviews Neuroscience*, 7(12):952–966, 2006. 9

R. Tomer, A.S. Denes, K. Tessmar-Raible, and D. Arendt. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5):800–809, 2010. 10, 14, 21, 39

R. Tomer, K. Khairy, F. Amat, and P.J. Keller. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nature Methods*, 9:755–763, 2012. 15

T.V. Truong and W. Supatto. Toward high-content/high-throughput imaging and analysis of embryonic morphogenesis. *Genesis*, 49(7):555–569, 2011. 2, 4

J.F.P. Ullmann, G. Cowin, N.D. Kurniawan, and S.P. Collin. A three-dimensional digital atlas of the zebrafish brain. *Neuroimage*, 51(1):76–82, 2010. 10, 16, 19, 66

C. Vonesch, F. Aguet, J.L. Vonesch, and M. Unser. The colored revolution of bioimaging. *IEEE Signal Processing Magazine*, 23(3):20–31, 2006. 2, 27

G. Walter, K. Büssow, A. Lueking, and J. Glökler. High-throughput protein arrays: prospects for molecular diagnostics. *Trends in Molecular Medicine*, 8 (6):250–253, 2002. 2

T. Walter, D.W. Shattuck, R. Baldock, M.E. Bastin, A.E. Carpenter, S. Duce, J. Ellenberg, A. Fraser, N. Hamilton, S. Pieper, et al. Visualization of image data from cells to organisms. *Nature Methods*, 7:S26–S41, 2010. 25

R.M. Warga and D.A. Kane. One-eyed pinhead regulates cell motility independent of squint/cyclops signaling. *Developmental Biology*, 261(2):391–411, 2003. 3

G.H. Weber, O. Rübel, M.Y. Huang, A.H. DePace, C.C. Fowlkes, S.V.E. Keränen, C.L.L. Hendriks, H. Hagen, D.W. Knowles, J. Malik, et al. Visual exploration of three-dimensional gene expression using physical views and linked abstract views. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:296–309, 2009. 25

M.C.M. Welten, S.B. De Haan, N. Van Den Boogert, J.N. Noordermeer, G.E.M. Lamers, H.P. Spaink, A.H. Meijer, and F.J. Verbeek. ZebraFISH: fluorescent in situ hybridization protocol and three-dimensional imaging of gene expression patterns. *Zebrafish*, 3(4):465–476, 2006. 15

R.P. Woods, S.T. Grafton, J.D.G. Watson, N.L. Sicotte, and J.C. Mazziotta. Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography*, 22(1):153–165, 1998. 21

R.P. Woods, M. Dapretto, N.L. Sicotte, A.W. Toga, and J.C. Mazziotta. Creation and use of a Talairach-compatible atlas for accurate, automated, nonlinear intersubject registration, and analysis of functional imaging data. *Human Brain Mapping*, 8(2-3):73–79, 1999. 12, 13

J.Q. Wu and T.D. Pollard. Counting cytokinesis proteins globally and locally in fission yeast. *Science*, 310(5746):310–314, 2005. 18

L. Yang, N.Y. Ho, R. Alshut, J. Legradi, C. Weiss, M. Reischl, R. Mikut, U. Liebel, F. Müller, and U. Strähle. Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. *Reproductive Toxicology*, 28 (2):245–253, 2009. 10

Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage*, 31(3):1116–1128, 2006. 70

C. Zanella, M. Campana, B. Rizzi, C. Melani, G. Sanguinetti, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarti. Cells segmentation from 3-D confocal images of early zebrafish embryogenesis. *IEEE Transactions on Image Processing*, 19(3):770–781, 2010. 19, 85

I. Zaslavsky, H. He, J. Tran, M.E. Martone, and A. Gupta. Integrating brain data spatially: spatial data infrastructure and atlas environment for online federation and analysis of brain images. In *Proc. Int. Workshop on Database and Expert Systems Applications*, pages 389–393. IEEE, 2004. 22

X. Zhou and P.D. Vize. Proximo-distal specialization of epithelial transport processes within the Xenopus pronephric kidney tubules. *Developmental Biology*, 271(2):322–338, 2004. 33

B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003. 20, 28

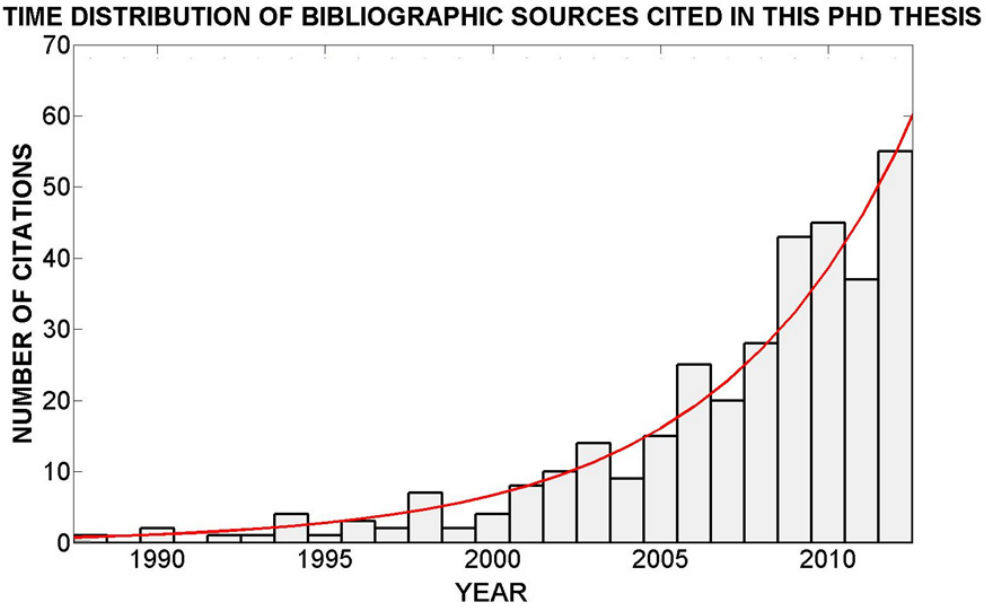

Figure: Histogram of the bibliographic sources cited in this PhD Thesis.
The temporal evolution in the number of citations concerning digital atlases can be modeled with an exponential law: $\boldsymbol{citations}=e^{(\lambda * year+\mu)}$, where μ is -348.7, λ is 0.1753 and the R^2 goodness of fit is 0.95.